PARKINSON'S DISEASE DETECTION

PROJECT PLANNING-CAPSTONE

TEAM MEMBERS

HARSHITH REDDY BODDIREDDY
PRANAY DATTA KAVUKUNTLA

Scope Statement and Objectives:

Building an extremely precise machine learning model utilizing the XGBoost algorithm is at the heart of the Parkinson's Disease Detection project's scope statement and goals. An accuracy goal of 90% or higher in identifying Parkinson's disease is the focus of this project's processing and analysis of speech recording data. This involves:

- Collecting a comprehensive dataset of vocal recordings from individuals diagnosed with Parkinson's disease and healthy controls.
- Data normalization and feature extraction are important steps in preprocessing for detecting Parkinson's disease symptoms including jitter, shimmer, and pitch changes. In order to prepare the data for the model, this pretreatment step involves building SSIS packages to transform the raw data into a staging table, and then utilizing SQL functions to refine the data for high-quality inputs.
- In order to train the model with high sensitivity and specificity in disease identification, we are using the XGBoost method on this dataset.
- Verifying the model's correctness and dependability in real-world situations by doing thorough testing on a different test set.
- Technical milestones include achieving a precision rate, recall rate, and F1 score that surpass current benchmarks in Parkinson's disease detection using vocal features.

Project Timeline/Schedule:

Task	Start Date	End Date
Project Kick-off	March 1, 2024	March 2, 2024
Data Collection	March 3, 2024	March 10, 2024
Data Preprocessing and Staging	March 11, 2024	March 17, 2024
Model Development (XGBoost Training)	March 18, 2024	April 1, 2024
Model Testing and Validation	April 2, 2024	April 15, 2024
Performance Evaluation	April 16, 2024	April 22, 2024
Final Review and Adjustments	April 23, 2024	April 27, 2024
Project Closure and Documentation	April 28, 2024	April 30, 2024

Task	Duration
Project Initialization	2 days
Data Collection	8 days
Data Preprocessing and Staging with SSIS	7 days

XGBoost Model Development	15 days
Model Testing & Evaluation	14 days
Performance Tuning	6 days
Documentation & Project Wrap-Up	4 days

Budget:

Budget Item	Allocation	Percentage	Details
Personnel Costs	$26,000	60%	Salaries for data scientists, developers, and a project manager for two months.
Software Licenses	$4,000	10%	Includes licenses for SQL Server, SSIS, and potentially commercial Python libraries.
Hardware and Cloud Services	$6,000	15%	Servers, GPUs for model training, and other cloud computing services as needed.
Data Acquisition	$4,000	10%	Purchase of vocal datasets, possible proprietary software for voice analysis.
Contingency Fund	$2,000	5%	Reserved for unforeseen costs, such as additional data or extended cloud service usage.
Total	42,000		

Requirements:

Data Requirements:

- A collection of voice samples, creating a strong and varied dataset for the model's training. Age, gender, and medical condition are examples of the kinds of metadata that should be attached with every collection.
- Streamlined procedures for collecting data to guarantee high-quality vocal recordings.

Software and Development Tools:

- Comprehensive licenses for SQL Server and SSIS for data staging and pre-processing tasks.
- Python environment setup with libraries for data analysis (pandas, NumPy), machine learning (scikit-learn, xgboost), and data visualization.
- Integrated development environment (IDE) like PyCharm or Visual Studio Code.

Hardware Requirements:

- High-performance servers with multi-core processors and high RAM capacities for efficient model training and data processing.
- Secure storage solutions with backup capabilities to ensure data integrity and availability.

Human Resources:

- Skilled data scientists experienced in ensemble learning techniques and feature extraction from audio data.
- A project manager with experience in agile methodologies to facilitate iterative development and timely delivery.
- Collaboration with medical professionals for domain expertise and validation of the model's diagnostic capabilities.

Compliance and Security:

- Compliance with global data protection regulations, ensuring that patient data is anonymized and handled with the utmost confidentiality.

Infrastructure and Operations:

- A well-defined code repository structure with branch management strategies to maintain code integrity throughout the development lifecycle.
- Continuous integration/continuous deployment (CI/CD) pipelines to automate testing and deployment processes.

Quality Criteria/Success Criteria:

1. **Model Accuracy**: The model must not only achieve the target accuracy of 90% but also maintain this performance level in cross-validation tests to mitigate overfitting.
2. **Precision and Recall Balance**: It's essential to maintain a balance between precision and recall, ideally with both metrics exceeding 90%, to ensure the model's reliability in distinguishing between positive and negative cases.
3. **F1 Score**: The F1 score, as a harmonic mean of precision and recall, should be maximized, reflecting a robust model that performs well in all aspects of classification.
4. **Model Response Time**: The classification response time should be optimized for clinical use, with a goal of processing and providing results within seconds from receiving a vocal sample.
5. **Usability**: The user interface and overall user experience should be intuitive and efficient, minimizing the time required for training and maximizing adoption rates among clinicians.
6. **Data Processing and Integrity**: The preprocessing and data handling pipeline must ensure the integrity of the data, with checks in place to detect and handle any anomalies or missing data points.
7. **Regulatory Compliance**: Compliance with healthcare regulations such as HIPAA in the US, GDPR in the EU, and other regional data protection laws is non-negotiable and must be verified before deployment.
8. **Scalability**: The infrastructure should be designed for easy scalability, capable of handling increasing data loads and user numbers without significant performance degradation.
9. **Maintenance and Support**: Post-deployment, the model should have minimal downtime and a clear plan for regular maintenance and updates based on the latest research findings and clinical feedback.

Project Resources:

Resource Category	Specific Resources	Quantity/Description	Estimated Cost
Human Resources	Data Scientists	2 (Expertise in ML, Data Preprocessing)	$16,000
	Project Manager	1 (Experienced in managing technology projects)	$8,000
	Domain Expert (Neurology)	1 (Insights into Parkinson's disease)	$2,000
Technical Resources	Servers/Cloud Computing	Cloud services for model training and data processing	$6,000
	Software Licenses	SSIS, SQL Server, Python IDEs, Machine Learning and Data Analysis Libraries	$4,000
Data Resources	Vocal Recording Database Access	Access to datasets of vocal recordings from patients and control groups	$2,000
Operational Resources	Office Supplies & Miscellaneous	Office materials, software for communication and collaboration	$1,000
	Collaboration Tools Subscription	Tools for project management, version control, and team communication	$500
Financial Resources	Contingency Fund	Budget reserved for unforeseen expenses	$2,000
Training Resources	Online Courses and Materials	For upskilling team members in the latest technologies and domain-specific knowledge	$500
			$42,000

Stake Holders List:

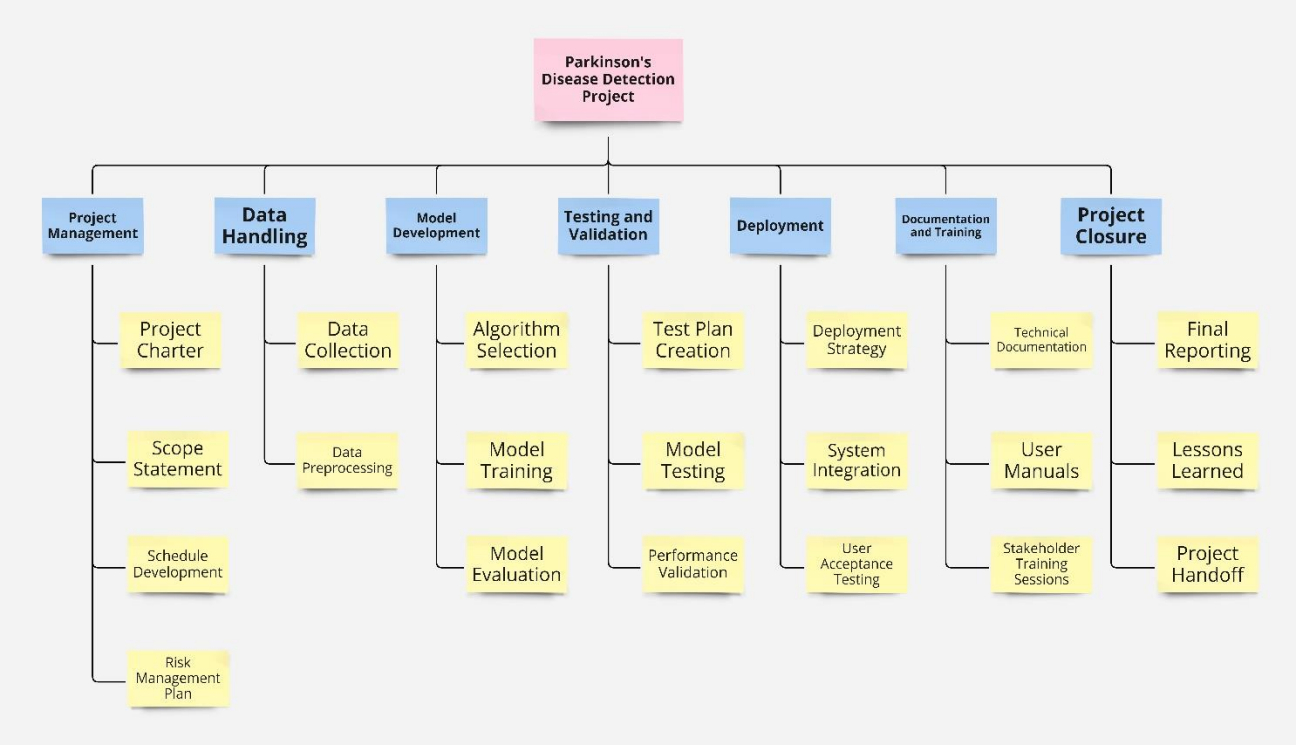
- **Project Team Members**: Comprising data scientists for model development, a project manager for overseeing project milestones, and technical support for infrastructure management.
- **Domain Experts**: Neurologists and other medical professionals specializing in Parkinson's disease, providing valuable insights into symptomatology and diagnostics.
- **Healthcare Providers**: Hospitals, clinics, and private practices that could implement the diagnostic tool in patient care, enhancing early detection and treatment strategies.

- **Funding Entities**: Government grants, private sector investments, and crowdfunding sources that provide the financial backing necessary for project development and scaling.
- **Technology Partners**: Companies that provide the essential software, hardware, and cloud computing resources necessary for data analysis and model development.
- **Research Community**: Academics and researchers in the fields of machine learning, medical informatics, and neurology, interested in the project's methods, results, and potential for further research.

Risk Management:

1. **Risk Identification**:
    - **Data Breaches**: The risk of unauthorized access to sensitive patient data.
    - **Inadequate Data Quality**: Insufficient, inaccurate, or biased data affecting model reliability.
    - **Technological Limitations**: Hardware or software constraints that could limit model performance or scalability.
2. **Risk Analysis**:
    - **Probability and Severity**:
        - Data breaches are considered high probability due to the increasing incidence of cyber-attacks and severe due to potential legal and reputational damage.
        - Inadequate data quality is moderate probability but high severity, as it directly impacts the model's diagnostic accuracy.
        - Technological limitations have a moderate probability and severity, depending on existing infrastructure and advancements in machine learning tools.
3. **Prioritization**:
    - Data security and data quality are top priorities due to their direct impact on project success and stakeholder trust.
    - Technological limitations are also critical but are considered manageable with current advancements.
4. **Risk Response Plans**:
    - Implement advanced cybersecurity measures, including encryption, access controls, and regular security audits, to protect against data breaches.
    - Establish rigorous data validation and cleaning protocols to ensure high data quality and model reliability.
    - Continuously evaluate and upgrade technological resources to meet project demands, ensuring model scalability and performance.

Work Breakdown Structure (WBS):



Responsibility Assignment Matrix (RAM):

Task or Deliverable	Project Manager	Data Scientists	Domain Expert	Technical Support
Project Charter	Lead	Consult	Consult	
Scope Statement	Lead	Consult	Consult	
Schedule Development	Lead	Support		
Budgeting	Lead			
Data Collection	Support	Lead	Consult	Support
Data Preprocessing		Lead	Support	Support
Model Development	Support	Lead	Support	Support
Model Testing		Lead	Support	Support
Deployment	Lead	Support		Lead
Documentation	Support	Lead	Consult	Support
Training Sessions	Support	Support	Lead	
Project Closure	Lead	Support	Support	

Here are the roles that are defined in this matrix:

- **Lead (L)**: Has full control over the task and is ultimately responsible for its completion.
- **Consult (C)**: Offers insight and perspective; facilitates conversation.
- **Support (S)**: Assists to the completion of the task; may be required to work under the supervision of the Lead.