NEXT
CAREER DEVELOPMENT ACADEMY

Graduation Project Proposal Form

1. Project Information

- **Project Title:** Healthcare Data Warehouse Integration
- **Course/Track:** Microsoft Data Engineer
- **Team Members:**

1. Khaled Ayman Farouk Ahmed
2. Metyas Monir Yousef Eskander
3. Mostafa Rabea Hashem
4. Noor Eldeen Mohammed Shrief
5. Mohamed Farouk Aboelfetouh

2. Project Overview

- **Objective:**

The objective of this project is to design and build a comprehensive data warehouse that integrates datasets from multiple diseases (diabetes, brain stroke, lung cancer, and heart problems). This warehouse will allow for efficient storage, management, and analysis of healthcare data, enabling advanced reporting, querying, and insights into patient demographics, health outcomes, and disease patterns. Additionally, the project will involve building machine learning models to predict whether a patient is at risk of having a particular disease

- **Scope of Work:**

1. **Data Extraction**:
   - Extract data from the provided CSV files containing disease-related information (diabetes, brain stroke, lung cancer, heart problems).
   - Perform data cleansing (removing duplicates, handling missing values, and correcting data formats) using tools like Azure Data Factory or Python scripts.
2. **Data Transformation**:
   - Transform the data into a consistent format across all datasets (e.g., standardizing columns like age, diagnosis, gender).
   - Feature engineering to prepare the dataset for machine learning, such as encoding categorical variables and normalizing numeric features.
3. **Data Warehousing**:
   - Design the data warehouse schema in Azure Synapse
   - Use a star schema, with fact table and dimension tables representing diseases, symptoms, and demographic data.
   - Create relationships between tables to enable complex queries for analytics
4. **ETL Pipeline**:
   - Use Azure Data Factory to create ETL pipelines that automate data extraction, transformation, and loading into the data warehouse
5. **Machine Learning**
   - Implement classification algorithms (Logistic Regression) to predict the likelihood of a patient having a specific disease.
   - Train and evaluate the models using metrics such as accuracy, precision, recall, and F1-score.

6. **Reporting & Visualization:**
   - Create interactive dashboards in Power BI for Analytics to provide insights into patient demographics, disease distribution, and prediction results

- **Expected Outcomes:**
  1. A well-organized data warehouse, integrating multiple disease-related datasets for analysis and machine learning.
  2. Fully automated ETL pipelines using Azure Data Factory, ensuring the warehouse is up-to-date with the latest patient data.
  3. Trained and deployed machine learning models that predict whether a patient is likely to have diabetes, brain stroke, lung cancer, or heart problems.
  4. A web service that can receive patient information and return disease predictions in real-time
  5. Interactive Power BI dashboards that visualize data trends, disease predictions, and model performance, providing valuable insights for healthcare providers.

Problem Statement

Healthcare providers face challenges in efficiently analyzing large amounts of patient data for early detection of diseases like diabetes, brain stroke, lung cancer, and heart problems. Manual methods are slow and lack scalability. This project aims to integrate diverse datasets into a centralized system, leveraging Azure and machine learning to predict disease risk in real-time. The solution will streamline disease prediction, support informed decision-making, and improve patient care outcomes.

3. Proposed Solution

- **Technologies Used:**
  1. Programming Language: Python (for data cleaning, feature engineering, and machine learning)
  2. Machine Learning: Scikit-learn (for model building and evaluation)
  3. Data Processing: Pandas, NumPy (for data manipulation)
  4. ETL Tools: SSIS or Python libraries (for extracting, transforming, and loading data)
  5. Azure : for pipeline and Azure Synapse for SQL query
- **System Architecture:**
  1. Data Ingestion: Transaction data is extracted from external sources and ingested into the MS SQL Server.
  2. Data Cleaning & Preprocessing: Data is cleaned and transformed using Python scripts before being used to train the model.
  3. Model Training: A disease detection model is built using historical transaction data, leveraging machine learning algorithms such as KNN
  4. Evaluation & Improvement: Regular monitoring of model performance to improve detection accuracy.

4. Resources Needed

- **Hardware/Software:**
  1. Python environment (Anaconda or any Python distribution)
  2. Python libraries: Pandas, NumPy, Seaborn, Matblotlib, Scikit-learn
  3. ETL tool: SSIS or custom Python scripts
  4. Computing infrastructure: Local machine or cloud-based servers for training and running models

5. Approval

- **Instructor/Advisor:** Moshira Ibrahim Ghaleb.

Signature: Moshira Ghaleb 1st Oct 2024