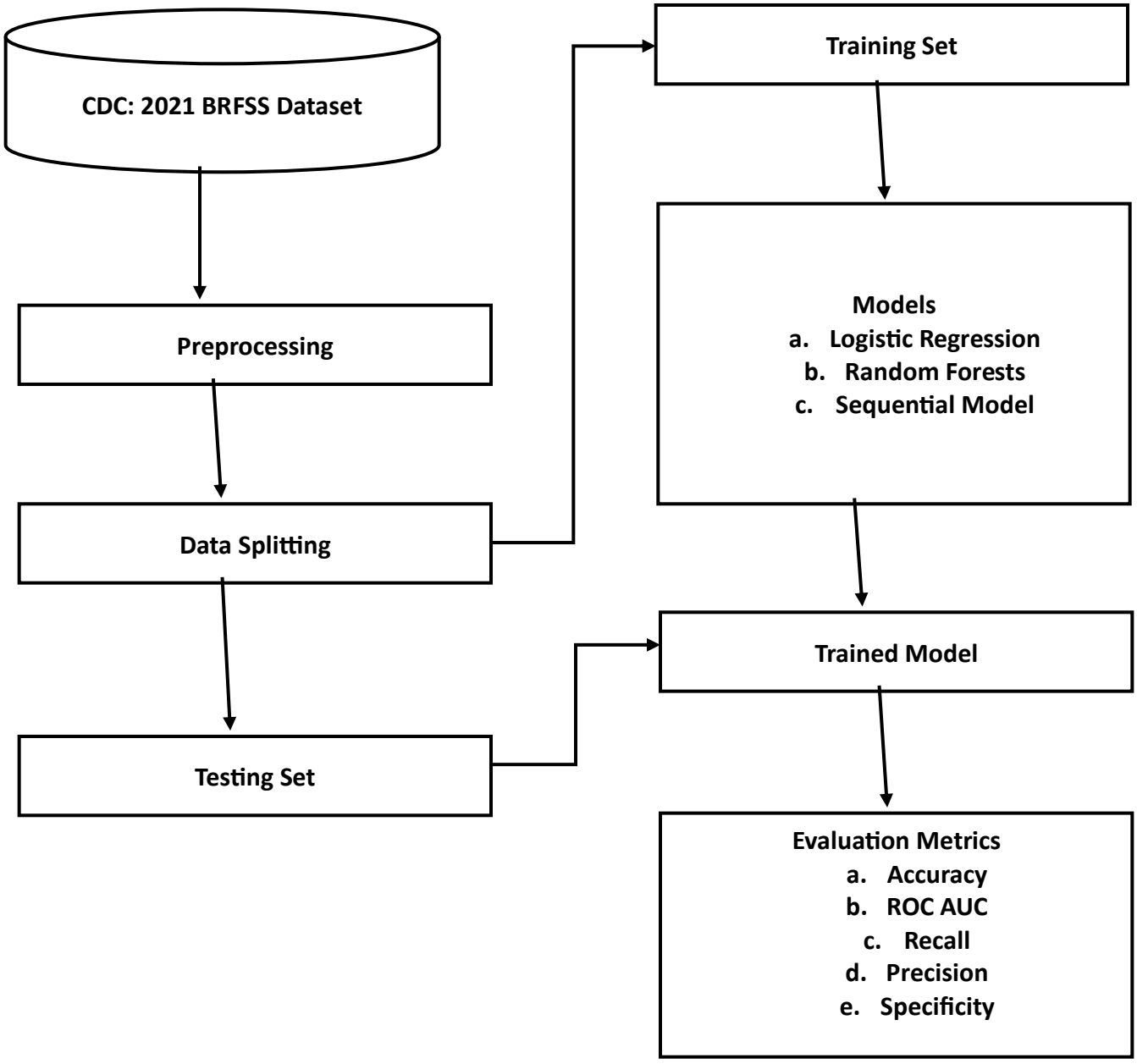
Diabetes Prediction Project ADD

Architectural Components Overview



Data Source: CDC – BRFSS 2021 Dataset

Technology Choice: Python's pandas

Data Conversion: SAS to CSV

Justification:
Python's pandas library is chosen for data manipulation and analysis tasks in this project. It provides a wide range of functionalities for handling and transforming data, making it suitable for working with the CDC - BRFSS 2021 Dataset. The choice of pandas is justified by its popularity in the data science community, extensive documentation, and ease of use.

Data Integration

Technology Choice: Not needed.

Justification:
Since the project only utilizes a single dataset, there is no need for additional technology for data integration. The dataset itself contains the required information for analysis, eliminating the need for data integration from multiple sources.

Data Repository

Technology Choice: Personal device

Justification:
The choice of using a personal device as a data repository is based on easy access and convenience. Storing the data on a personal device allows for quick retrieval and analysis without relying on external systems or network connectivity.

Discovery and Exploration

Technology Choice: Python's libraries: pandas, numpy, seaborn, matplotlib

Justification:
Python's pandas, numpy, seaborn, and matplotlib libraries are chosen for data exploration and transformation tasks. Pandas provides powerful data manipulation capabilities, numpy offers efficient numerical operations, seaborn enables visually appealing data visualization, and matplotlib allows for comprehensive plotting capabilities. These libraries together provide a comprehensive toolkit for exploring and transforming the data. The cleaned data is stored on the personal device to maintain control and easy accessibility during the analysis phase.

Actionable Insights

Technology Choice: Python's libraries: matplotlib, sklearn, and tensorflow

Justification:
Python's libraries, including matplotlib, sklearn, and tensorflow, are selected for generating actionable insights from the analyzed data. Matplotlib allows for visualizing the results, sklearn provides machine learning algorithms for predictive modeling, and tensorflow offers a versatile framework for building and deploying machine learning models. These libraries are widely used in the data science community and provide ample functionality for deriving insights and building predictive models.

Applications / Data Products

Technology Choice: Jupyter Notebook

Justification:
Jupyter Notebook is chosen as the technology for developing applications or data products. Jupyter Notebook is an interactive environment that allows for combining code, visualizations, and explanatory text, making it suitable for documenting the project's analysis, results, and conclusions. Its flexibility and interactivity facilitate collaboration and presentation of the project's findings.

Security, Information Governance and Systems Management

Technology Choice: Not needed.

Justification:
No specific technology is needed for security, information governance, and systems management in this project. The BRFSS dataset used is publicly available and do not contain sensitive information. Therefore, additional security measures or systems management tools are not required.