**PREDICTING HEART DISEASE**

**USING AWS AND PYSPARK**

**Project Group Number: 22**

|  |  |
| --- | --- |
| **Name** | **ID Number** |
|  |  |
|  |  |
|  |  |
|  |  |

**Table of Contents**

|  |  |
| --- | --- |
| **Content** | **Page Number** |
| Objects & Goals | 3 |
| Motivation | 3 |
| Significance | 3 |
| Project and Milestones | 4 |
| Visualization Diagram | 5 |
| Objectives | 6 |
| Features | 6 |
| Conclusion | 7 |
| References | 7 |

**Objects & Goals**

* The identification and precise forecasting of cardiovascular disease are of utmost importance in minimizing its consequences, decreasing death rates, and enhancing the general well-being of those afflicted. In the current epoch of data-centric healthcare, the amalgamation of advanced technologies like Amazon Web Services (AWS), Hive and PySpark offers a propitious pathway for the advancement of prognostic models where hive offers the data analysis by using HiveQL which is similar to SQL.

**Motivation:**

* Despite its prevalence rising as a result of contemporary lifestyles and dietary practices, heart disease continues to be one of the top causes of mortality globally. To lessen the effects of and save lives from heart disease, early identification and prediction are essential. In this project, we suggest using the strength of AWS (Amazon Web Services), Hive is to analyze the data and PySpark to create a heart disease predicting model, which may help medical practitioners make prompt and accurate diagnosis.

**Significance:**

* **Early identification:** It is crucial for successful treatments since heart disease often has modest symptoms. Our initiative attempts to provide a prediction tool that may identify vulnerable people even before they exhibit outward signs (Kolli et al. 2022).
* **Resource Efficiency:** It can analyze huge datasets quickly using AWS, Hive and PySpark, which makes our prediction model scalable and appropriate for healthcare organizations with a range of patient loads.
* **Predictive models:** It may help in personalized medicine by helping to customize treatment regimens for specific patients, improving the efficiency of therapies, and reducing adverse effects.

**Project Milestones or deliverables:**

* Planning a project, forming a team, getting to know each other, identifying the previous existing models, and establishing the tasks and timeline of the project.
* Finding the correct dataset related to healthcare and heart disease and to loading that data into AWS S3, which is the most secure and durable,
* To analyze the dataset in Hive using HiveQL by implementing the queries on the huge data.
* By using PySpark to integrate with AWS to preprocess the data, which is for data cleaning and data transformation.
* Identifying the correct machine learning algorithm for splitting up the data, implementing it in PySpark, and deploying it in AWS
* Conducting the entire model or entire code for end-to-end testing without any errors.
* Executing the code and displaying the output and visualization diagrams for better understanding and to show accurate results.
* Lastly, documenting the entire model and project presentation as per the guidelines given by the professor.

**Visualization Diagram:**

A diagram of a project

Description automatically generated

**Objectives:**

* To ascertain the most relevant attributes and maybe generate novel ones that enhance the accuracy of cardiac disease prognosis. This endeavor requires a combination of expertise in the field and data-driven methodologies.
* To analyze the data, we will use Hive. It is mainly used to query and analyze the large amount of dataset by using HiveQL language to get accurate results.
* To construct and train machine learning models using PySpark. A range of methods, including logistic regression, random forests, and gradient boosting, will be used to conduct experiments and identify the most effective model.
* To determine the prediction model performance, it is recommended to use suitable measures such as precision, F1-score, accuracy, recall, and ROC-AUC. Cross-validation will be employed to guarantee the robustness of the model.
* To use the capabilities of AWS, it is recommended to avail several services such as EC2 for computational resources, S3 for data storage, and Sage Maker for deploying models. This approach guarantees scalability, security, and cost-efficiency (Ali et al. 2022).

**Features:**

* **Data Variety:** To improve prediction accuracy, the study will make use of a variety of characteristics, including genetic data, medical history, lifestyle variables (such as food and exercise), and demographic data.
* **Scalability:** Respect to AWS services, the predictive model will be able to manage enormous datasets and change with the needs of the healthcare industry (Mohan et al. 2019).
* **Real-time Prediction:** The created model will provide real-time forecasts, allowing for prompt treatments and lowering the risk of consequences from heart disease.
* **Interpretability**: We'll work to make the model understandable so that medical practitioners can comprehend the elements that go into each prediction.

**Conclusion:**

* In conclusion, addresses the crucial requirement for heart disease early detection. We intend to build a scalable, effective, and precise prediction model by using Hive, PySpark's distributed computing capabilities and AWS's cloud architecture. By supporting healthcare professionals in identifying persons at risk and customizing therapies for better results, this initiative has the potential to have a substantial influence on public health.

**References**

1. Ali, F., El-Sappagh, S., Islam, S.R., Kwak, D., Ali, A., Imran, M. and Kwak, K.S., 2020. A smart healthcare monitoring system for heart disease prediction based on ensemble deep learning and feature fusion. Information Fusion, 63, pp.208-222.
2. Kolli, V.K., 2022. Prediction of Cardiovascular Risk Using Extreme Learning Machine-Tree Classifier on Apache Spark Cluster. Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science), 15(3), pp.443-455.
3. Mohan, S., Thirumalai, C. and Srivastava, G., 2019. Effective heart disease prediction using hybrid machine learning techniques. IEEE access, 7, pp.81542-81554.