Deep Dive on Prevalence of Smoking and Heart Diseases in the US

Identifying focal reasons for heart related issues with big data analysis

By

Alisha Minj

WG56858

Exploring the Big Data Problem in Healthcare

Lack of data infrastructure to capture and store large amounts of data

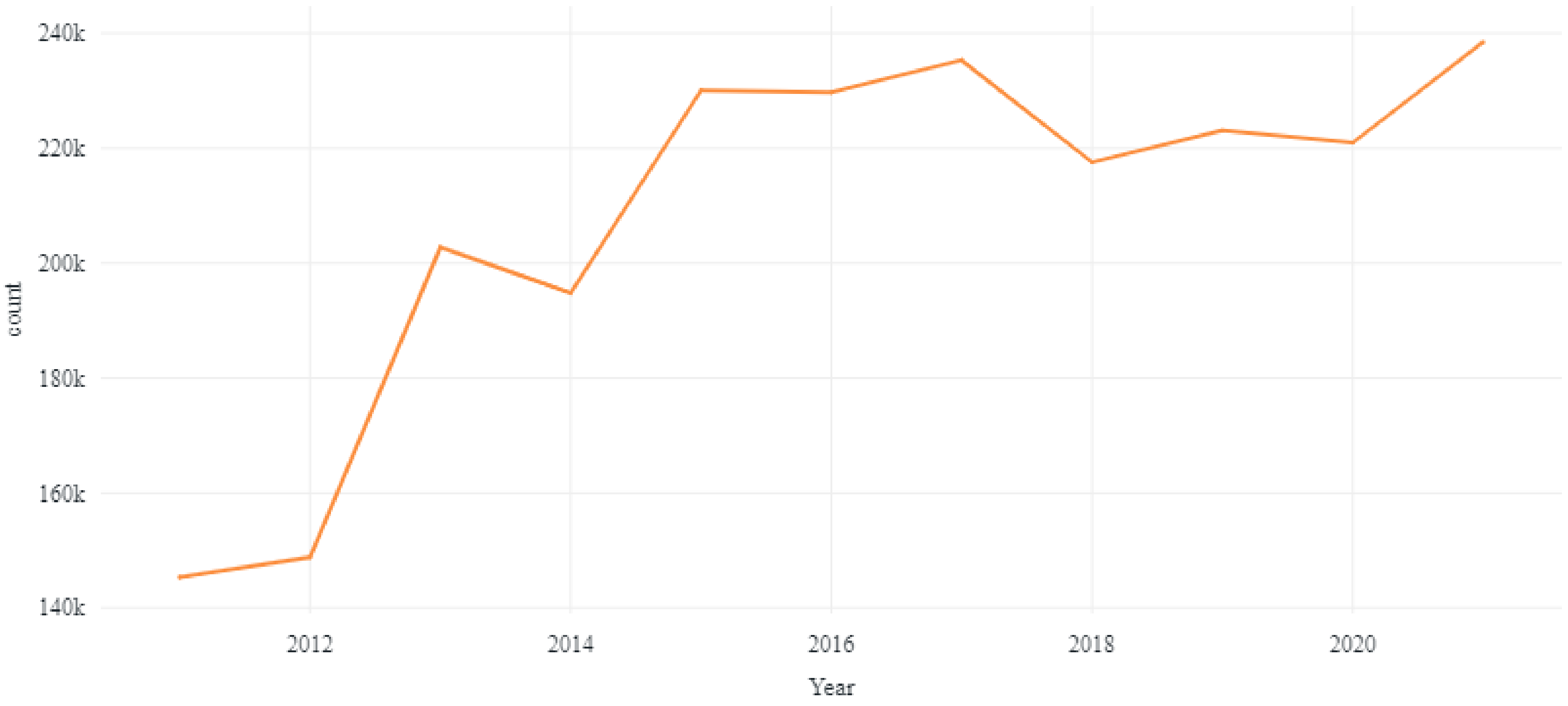
Inadequate analytics capabilities to analyze and interpret data.

Inability to effectively aggregate data from multiple sources.

Ensure data security and Comply with regulations

Project BRFSS

Behavioral Risk Factor Surveillance System

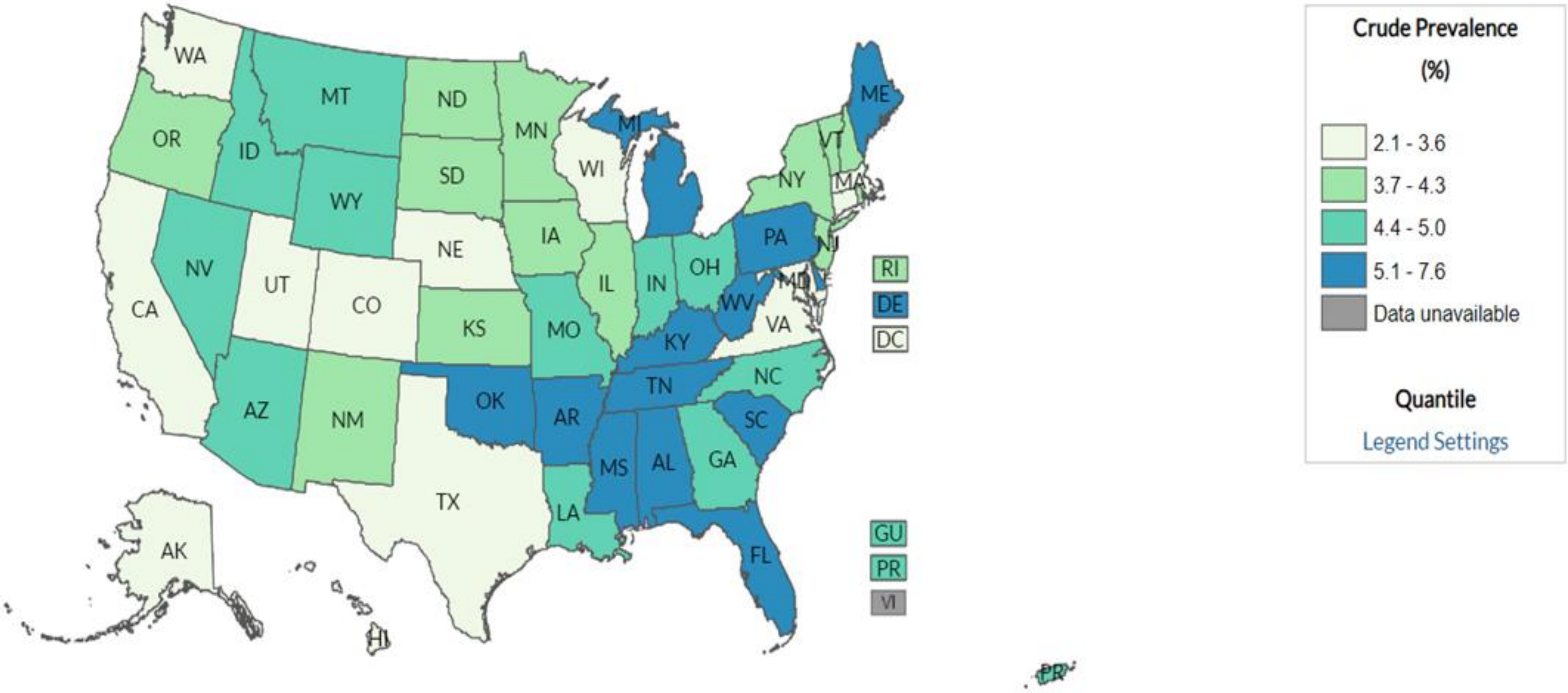


https://github.com/alishaminj12/Final_Project

Interpret

- To identify habits of patients prone to Smoking
- Identify prevalence of heart disease

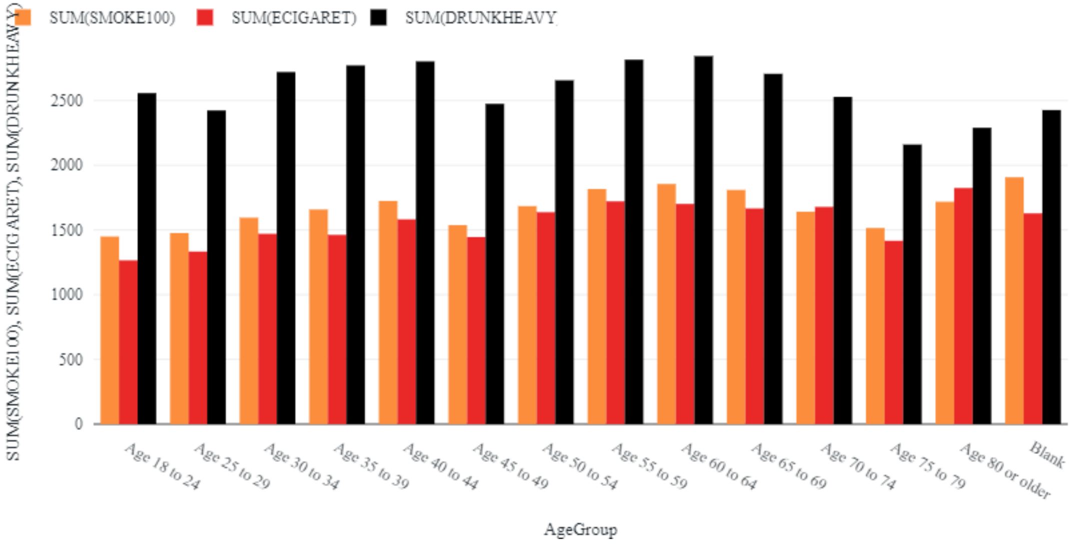
Ever told you had a heart attack (myocardial infarction) ?

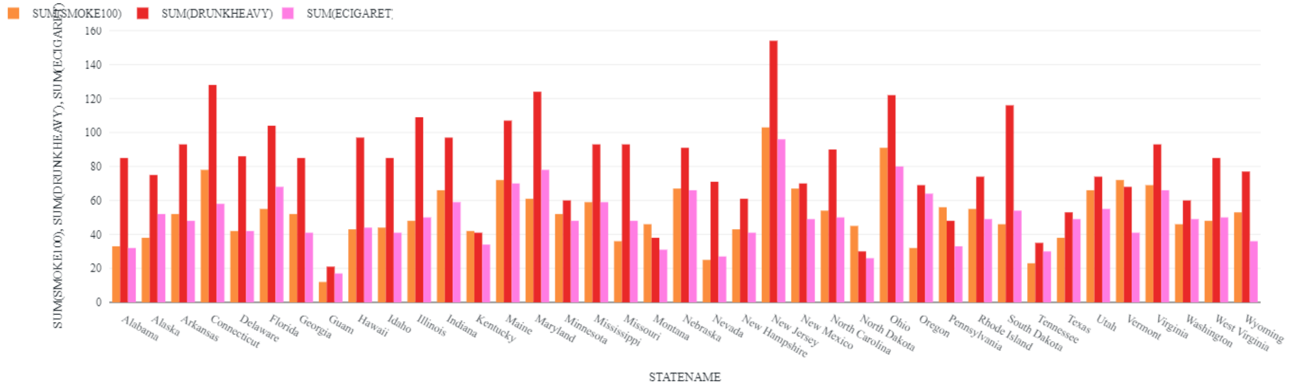


Analysis

Smoking Patterns

- Most of the respondents prefer **drinking**
- In **New Jersey**, has maximum Alcohol intake, smoking and E-cigarettes are also popular
- **Acceptance** of E- cigarettes

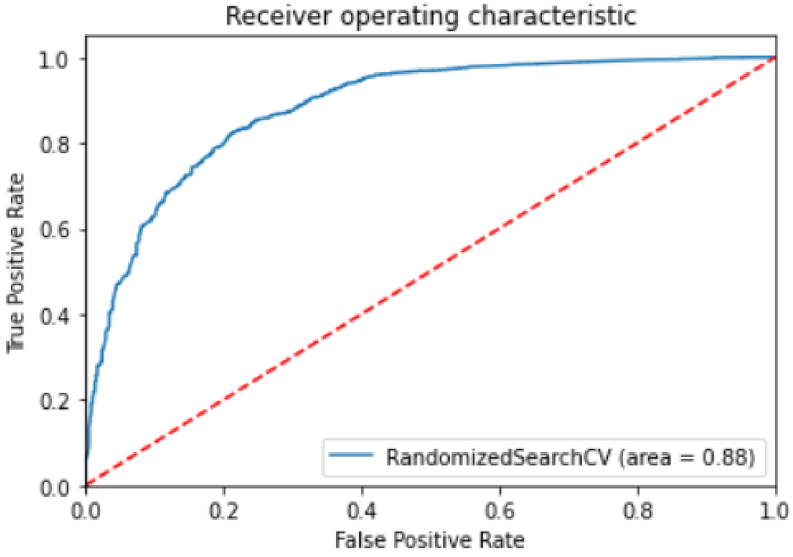


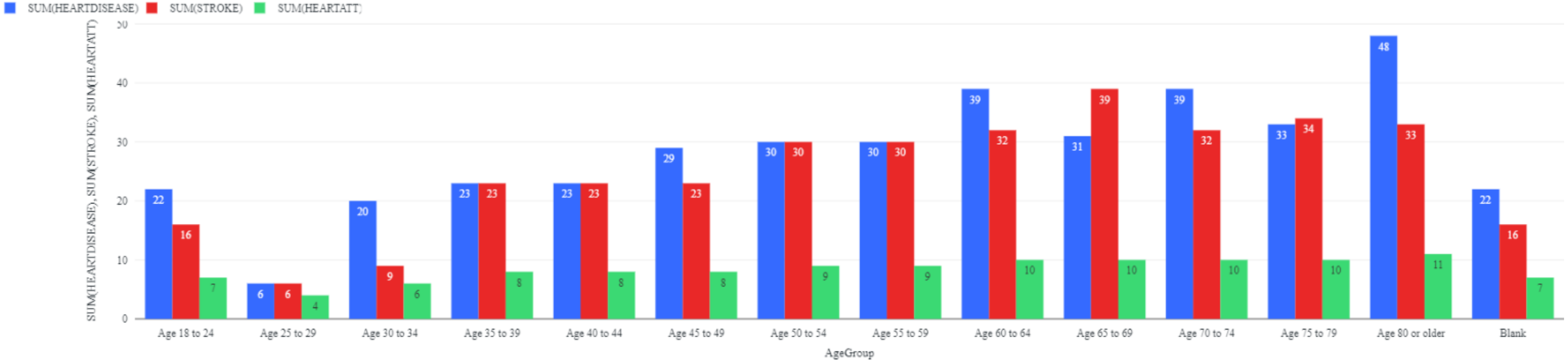


Analysis

Heart Attack Prediction

- Ohio, Maine, Washington, Virginia are having more prone heart attacks
- Performed logistic Regression



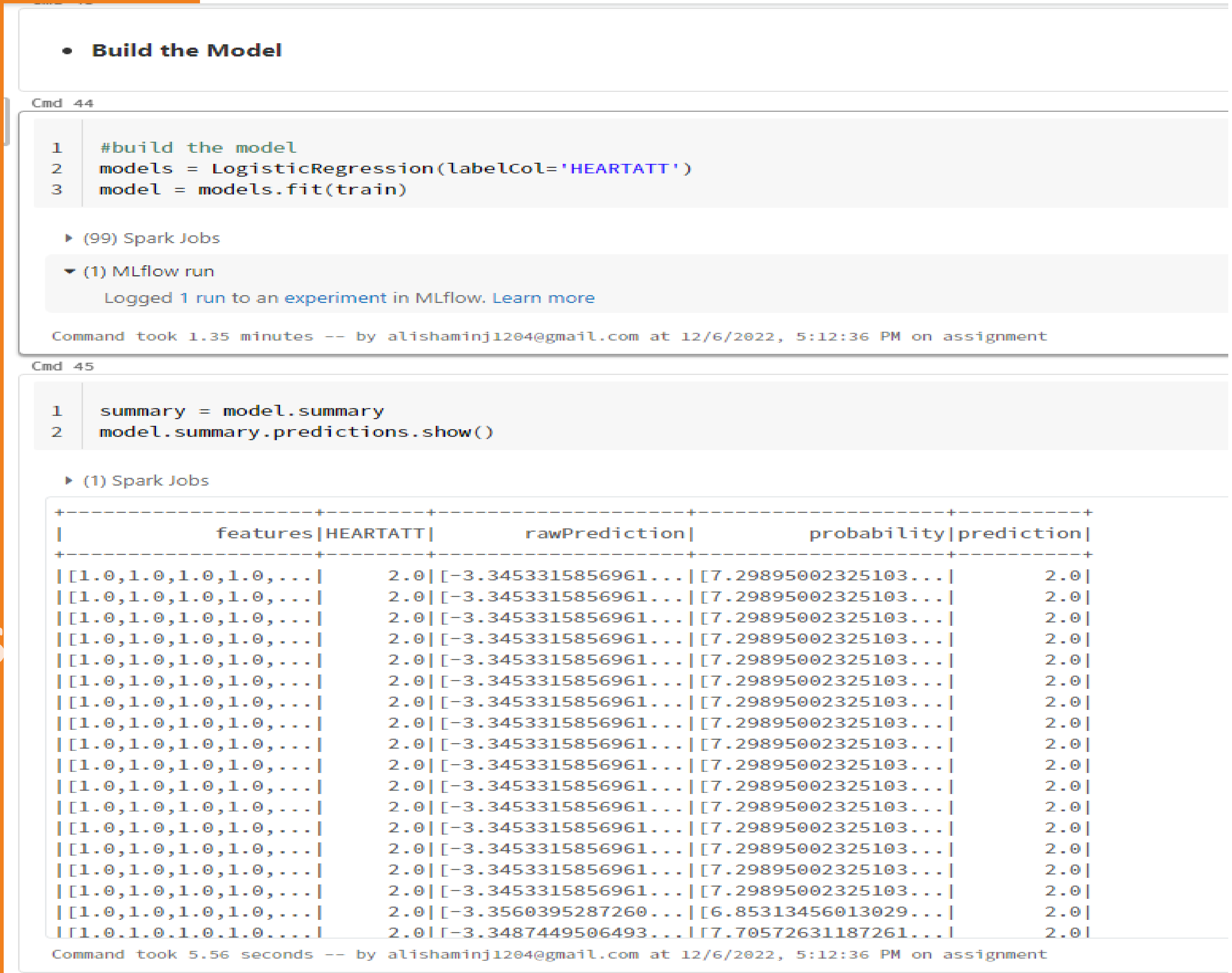


Working on

markdown in

PySpark

Databricks



- **Build the Model**

Cmd 44

```
#build the model
models = LogisticRegression(labelCol='HEARTATT')
model = models.fit(train)
```

▸ (99) Spark Jobs

▾ (1) MLflow run

Logged 1 run to an experiment in MLflow. Learn more

Command took 1.35 minutes -- by alishaminj1204@gmail.com at 12/6/2022, 5:12:36 PM on assignment

Cmd 45

```
summary = model.summary
model.summary.predictions.show()
```

▸ (1) Spark Jobs

```
+--------------------+--------+--------------------+--------------------+----------+
|            features|HEARTATT|       rawPrediction|         probability|prediction|
+--------------------+--------+--------------------+--------------------+----------+
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3453315856961...|[7.29895002325103...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3560395287260...|[6.85313456013029...|       2.0|
|[1.0,1.0,1.0,1.0,...|     2.0|[-3.3487449506493...|[7.70572631187261...|       2.0|
```

Command took 5.56 seconds -- by alishaminj1204@gmail.com at 12/6/2022, 5:12:36 PM on assignment

Recommendation

❑ Include specific question and take of the data inconsistencies.

❑ Develop a strategy for data storage and management.

❑ Employ data security and privacy measures to ensure the safety of patient data

THANK YOU!