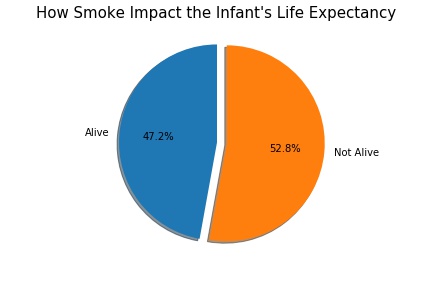
|  |  |
| --- | --- |
| **Soal Kualifikasi**  *Qualification Case* |  |
| COMP6579  Big Data Processing |
| **Post-Training**  *Case created on Odd 2019/2020* | **COMP6579-AW** |

**Budi Analytics**

Budi Analytics is one of the biggest Data Analytics companies in Zimbabwe. Recently, they are tasked from National Center for Health Statistics to **build a predictive model** to predict **infant chances of surviving** based on the data they have provided in “**births.csv**”. Sudut, the CEO of Budi Analytics, is asking you as one of the data scientists in the company to solve the problem given by them.

**Section 1. Data Visualization**

Your first task is to find and understand the **outliers** and **patterns** in data by using visual elements like charts, graphs, and maps. You are **free to use any data visualization tools** available in **anaconda smokes distribution**. Show **at least THREE** different visualizations (with different types) and don’t forget to write the **DESCRIPTION** or **CONCLUSION** about the data you plotting to make it more understandable. Below is one of the example:



**Conclusion**: We can see from the pie chart above that **smoke will affect the infant’s life**. For the mother who doesn’t smoke, the percentage for the infant’s to live is around **5.6% more** than the mother who smoke.

**Section 2. Building a Predictive Model**

After you find the patterns in the data, you can use the analysis to help you build the predictive model. You are **free to use any methods** (Clustering, Classification, or Regression) and algorithms that are available in the **MLlib**. If you are using more than one method, pick a model with the **best accuracy** and don’t forget to **write the steps required** to build the model (including preprocessing, recoding, etc). It’s important to notes that the main objective of building the predictive model is to **predict infant chances of surviving** (whether an infant is alive or not alive). Below are the steps:

1. **Load and Transform Data**

You need to load all raw data available that will be trained later. After that, you can transform the data until it is suitable for you to use them for training, although this is **optional**. For example, you may recode the INFANT\_ALIVE\_AT\_REPORT variable’s value to be either 1 or 0.

1. **Data Preprocessing**

Don’t forget to **handle** (remove, replace, etc) any **outliers** and **missing values** in the data. You may use any preprocessing technique available such as normalization, mapping values, etc.

1. **Select Features**

After preprocessing the data, you need to **select some features** available in the data to use it for training. Once again, you are free to use any feature selection algorithm or methods available in the lab. Simple analysis from data visualization also accepted.

1. **Preparing Data**

In this phase, you need to **prepare all the data for training**. For example, if you are using classification algorithm, you may create RDD of LabeledPoints to be used as input for the machine learning algorithm.

1. **Predicting Input**

After the model is generated, you can **test the model** to predict whether the infant chance of surviving

1. **Model Evaluation (Optional)**

**Files to be collected**:

* .ipynb file
* Data visualization and model documentation (free format, ex: docx)
* Any supporting material (optional)

**Notes:**

* You **don’t** need to upload any datasets or file from the case
* Please use **logical analysis** to solve the problem presented in the case and **don’t** use any external data or assumption that are not included in the case
* You can find the documentation of the data using User Manual provided by them in “**manual.pdf**”

**Data References:**

* <https://www.cdc.gov/nchs/>