Predicting Drinkers Using Body Signal Data

HARITH PALAKAL MAVILAVALAPIL

ABSTRACT

This project aims to leverage big data analytics to explore the drinking habits of individuals and predict whether someone is likely to be a drinker. Using a comprehensive dataset from Kaggle containing information on various biological factors, we navigated through the challenge of high dimensionality by focusing on a curated set of key features. Through meticulous data cleaning, preprocessing, and exploratory data analysis, we prepared these attributes for subsequent modelling efforts. This study unveiled the efficacy of machine learning algorithms like logistic regression and support vector machines in predicting an individual's propensity for drinking. Additionally, we underscored the importance of feature selection in handling datasets with numerous dimensions. Furthermore, we employed Tableau for data visualization, enabling us to discern features closely associated with our target variable. By plotting insightful figures, we gained a deeper understanding of the variables and their correlation with the target variable. This abstract encapsulates our endeavour to glean actionable insights into drinking behaviour using a blend of big data analytics and visualization techniques.

Keywords: smoking, drinking dataset, logistic regression, support vector machine, feature

selection

**CONTENT**

Introduction………………………………………………………………………………….4

Background/Related Work………………………………………………………………. .5

Dataset Analysis …………………………………………………………………………….5

DATA SET OVER VIEW……………………………………………………………………….6

Data preparation………………………………………………………………………………7

Methodology……………………………………………………………………………….....6

Logistic Regression…………………………………………………………………8

Support Vector Machine (SVM)……………………………………………….......8

Software Installation………………………………….……………………………………8

Experimental Section…………………………………………………………………………7

Tableau Visualization………………………………………………………………………..10

Result Discussion……………………………………………………………………………..11

Conclusion………………………………………………………………………………….….11

Future Works……………………………………………………………………………….....12

Social and Ethical Impact …………………………………………………………….….….12

References ………………………………………………………………………………....…..13

Appendix………………………………………………………………………………………...14

1. INTRODUCTION

In modern society, comprehending human behaviour holds significant importance across various domains, from public health endeavours to tailored marketing strategies. Of particular interest is the study of drinking habits, given their profound impact on individual well-being and societal dynamics. This research leverages the capabilities of big data analytics to delve into the intricacies of drinking behaviour, aiming to predict whether individuals are likely to consume alcohol. By utilizing a comprehensive dataset sourced from Kaggle, which encompasses a wide array of biological factors, we embark on a journey to uncover patterns and predictors associated with drinking behaviour.

Despite the dataset's richness, its high dimensionality poses a formidable challenge, necessitating a strategic approach to feature selection and model development. Therefore, our project places emphasis on curating a focused set of key features through rigorous data cleaning, preprocessing, and exploratory data analysis. By distilling the essence of the dataset into a manageable subset of pertinent attributes, we lay the groundwork for effective modelling and prediction of drinking tendencies.

Central to our investigation is the application of machine learning algorithms, including logistic regression and support vector machines, to discern patterns in drinking behaviour. Implemented using PySpark, these algorithms undergo thorough training and evaluation to uncover the most significant factors influencing an individual's likelihood of being a drinker. Moreover, we highlight the critical role of feature selection in navigating datasets characterized by high dimensionality, ensuring robust and interpretable model outcomes.

Augmenting our analytical efforts is the utilization of Tableau, a powerful data visualization tool. Through Tableau's capabilities (Batt et al., 2020), we elucidate relationships between the target variable and various features, providing intuitive insights into the drivers of drinking behaviour. By presenting visually engaging figures and interactive dashboards, we unravel the complexities surrounding alcohol consumption, empowering stakeholders with actionable insights for decision-making.

In summary, this study adopts a comprehensive approach to understanding drinking habits, integrating the analytical prowess of big data analytics with the interpretive clarity of data visualization. By untangling the intricate interplay of biological factors and drinking behaviour, our aim is to equip stakeholders with actionable insights that can inform interventions in public health, targeted marketing initiatives, and broader societal endeavours.

1. **Background/ Related Work:**

Data-Driven Approaches to Prediction of drinker or not: The advent of big data and computational methodologies has revolutionized the field of identifying individuals with drinking habits. Traditional qualitative methods are being supplanted by data-driven approaches, particularly in the context of machine learning (Domingos, 2012). By harnessing advanced algorithms, researchers can now analyse extensive datasets to uncover significant patterns and insights into drinking behaviour. This enables the development of predictive models capable of determining whether an individual is prone to drinking or not.

1. **DATA ANALYSIS:**

In this project, we use a dataset that includes several variables connected with various biological processes. Our analysis is inspired by the methodology and discoveries described in the previously mentioned related work

Correlation analysis has traditionally been utilized to identify patterns and relationships among various lifestyle choices. While these methodologies are effective in revealing associations, they do not inherently offer a means to predict outcomes based on input variables. This gap underscores the significance of regression analysis, which becomes indispensable for predicting outcomes. In this project, we leverage regression analysis to conduct our classification, allowing us to quantify the impact of drinking habits. Additionally, we incorporate a support vector machine algorithm to compare the accuracy of our predictive models.

|  |  |
| --- | --- |
| DATA SET OVER VIEW | |
| **DATA SOURCE** | <https://www.kaggle.com/datasets/sooyoungher/smoking-drinking-dataset> |
| **DATASET DESCRIPTION** | Smoking, Drinking Dataset, sourced from Kaggle which Predict smokers and drinkers using body signal data This dataset consists of 991346 persons body signal data |
| **NAME** | For this project, the Data Frame (**df**) will be referred to as data. |

|  |  |  |
| --- | --- | --- |
| **NO** | **Description** | **Type** |
| 1 | Age | integer |
| 2 | Sex | string |
| 3 | Height | integer |
| 4 | Weight | integer |
| 5 | waistline | Float |
| 6 | Sight left | Float |
| 7 | Sight Right | Float |
| 8 | Hear left | Float |
| 9 | Hear right | Float |
| 10 | SBP  Systolic blood pressure[mmHg] | Float |
| 11 | DBP  Diastolic blood pressure[mmHg] | Float |
| 12 | BLDS  BLDS or FSG(fasting blood glucose)[mg/dL] | Float |
| 13 | Tot\_chole  total cholesterol[mg/dL] | Float |
| 14 | HDL\_chole  HDL cholesterol[mg/dL] | Float |
| 15 | LDL\_chole  LDL cholesterol[mg/dL] | Float |
| 16 | Triglyceride  triglyceride[mg/dL] | Float |
| 17 | Hemoglobin  hemoglobin[g/dL] | Float |
| 18 | urine\_protein | Float |
| 19 | serum\_creatinine | Float |
| 20 | SGOT\_ASTSGOT(Glutamate-oxaloacetate transaminase) AST(Aspartate transaminase)[IU/L] | Float |
| 21 | SGOT\_ALT  ALT(Alanine transaminase)[IU/L] | Float |
| 22 | gamma\_GTP | Float |
| 23 | SMK\_stat\_type\_cd  Smoking state, 1(never), 2(used to smoke but quit), 3(still smoke) | Float |
| 24 | DRK\_YN  (Drinker or Not) | string |

TABLE 1. DATASET FEATURES

The dataset, sourced from Kaggle, comprises 991,346 instances, each representing an individual's body signal data. It consists of 24 attributes, with no missing values observed. 495488 people drink, while 495858 people do not drink. These attributes encompass 2 categorical values, 3 integers, and 19 float values. The final attribute serves as our target variable, denoting whether the individual is classified as a drinker or non-drinker, indicated by string values. All attributes are some body signals, in this project by analysing those attributes we are trying to making a machine learning model with good accuracy.

1. **Data preparation**:

Data Cleaning**:** Data cleaning is an important phase in the data preparation process that involves identifying and addressing missing values, outliers, and inconsistencies in the information. Checked missing value and found that there is no missing value in the data set.

Feature Encoding**:** For better performance we have to convert categorical value in to binary values, in this project we encoded categorical value in to numerical value by using StringIndexer method(*StringIndexer — PySpark Master Documentation*, n.d). Checked the data to find any null values.

1. **Methodology**:
2. Logistic Regression:

Logistic Regression is a statistical approach used to analyse datasets in which one or more independent variables influence a binary outcome. This method is best effective when the result variable is dichotomous, which means it has only two possible values. We utilise logistic regression to predict a binary outcome, such as Yes/No, based on a set of independent variables (Jr et al., 2013).

1. Support Vector Machine:

The Support Vector Machine (SVM) is a strong classification algorithm that identifies the best hyperplane to separate classes by maximising the margin. It uses kernel functions to handle both linear and nonlinear data, ensuring robust classification. Regularisation parameters regulate overfitting. SVM's efficiency in binary and multi-class classification tasks is well known (Cortes & Vapnik, 1995). It remains a popular choice for numerous machine learning applications because of its adaptability and performance.

1. Software Installation:

As required by the coursework, the installation of the Java, Hadoop and Spark was done on Ubuntu after installing a virtual machine on my computer. The java software was installed using the ‘sudo apt install’ command, I also got the required versions of Hadoop and spark. Java version – 1.8.0\_392, Hadoop-2,7.3 and spark -2.3.0 has been installed. Installation proof is attached on appendix on fig1, fig2,2.a,2.b,2.c and fig3.a,3.b respectively.

1. **EXPERIMENTAL SECTION**

The first step involved the installation of Java, Spark (Fig-4), and PySpark on Google Colab (Fig-5). Subsequently, essential libraries, including those for machine learning from PySpark, were imported. After setting up the environment and loading the libraries, the dataset, Smoking\_Drinking.csv, was imported. A SparkSession was created using SparkSession.builder to interact with Spark in Python (Fig-6). Then, the dataset was loaded into a Spark DataFrame named 'data'. Null values were checked (Fig-7), and the initial DataFrame information, schema, and the first few rows were displayed for inspection (Fig-8).

The subsequent step involves preprocessing the categorical columns in the dataset to prepare them for machine learning algorithms (Fig-9). Initially, it identifies the categorical columns necessitating transformation. Employing StringIndexer, each distinct string value within these columns is mapped to a unique numerical index, rendering them suitable for utilization with machine learning models.

VectorAssembler transformation to create a new DataFrame output containing the original columns along with the newly assembled feature vector column. Then Displayed the first 3 rows of the "features" column from the DataFrame output(Fig-10).

Created new DataFrame named data\_final by adding all features and our target variable "DRK\_YN\_encoded" for model creation (Fig-11). The next stage is data splitting, in which I will divide the data into training and test sets. Approximately 70% of the data will be given to the training set, with the remaining 30% going to testing. Using seed=42 ensures that the data is split consistently each time. This improves consistency across multiple runs (Fig 12).

The next stage is to start the Model Training and Performance evaluation using ROC AUC, F1 score and Precision. Marzban (2004) investigates the ROC curve and its area under the curve (AUC) as critical performance indicators, particularly for measuring the effectiveness of forecast models. We created 2 models for this project.

1. First, a **Logistic Regression model** is created and fitted with the training data. A summary table is generated to evaluate the model. The model's performance is evaluated using Binary Classification Evaluator.
2. To compare models, a **Support Vector Machine model** is created. After fitting the model with the data, a summary is printed for observation. The model is evaluated using ROC AUC.
3. **TABLEAU VISUALIZATION**

Fig 18 shows the total number of males and females present in this data set. Visualization in Fig-19 depicts the relationship between age and drinking habits based on the dataset. Notably, individuals aged between 20 and 40 exhibit a higher propensity for alcohol consumption compared to non-drinkers. This suggests a prevalent trend of alcohol initiation during the 20s and 30s. Moreover, a decline in the number of drinkers is observed beyond the age of 50

Figure 20 illustrates the relationship between gender and drinking behaviour, revealing a notable disparity between men and women. According to our dataset, a higher proportion of men are identified as drinkers compared to women. Conversely, the majority of women in the dataset are categorized as non-drinkers, highlighting a gender-based discrepancy in alcohol consumption habits.

Figure 21 depicts the relationship between drinking and smoking habits. Upon analysis, it is observed that individuals who do not smoke are more likely to be non-drinkers. This inference is drawn from the significantly higher number of non-smokers who do not consume alcohol compared to those who do. Conversely, individuals who still smoke are more inclined to be drinkers. The bar plot suggests a modest correlation between smoking and drinking behaviours.

Figure 22 illustrates a Tree map depicting the relationship between SBP (Systolic Blood Pressure) and drinking behaviour (Drinker Y/N). Upon analysis, it is observed that individuals with higher SBP counts are more likely to be drinkers.

To explore any potential patterns between BLDS (Blood Sugar) or FSG (Fasting Blood Glucose) and our target variable, Fig-23 was generated. However, the highlighted table and packed bubble chart indicate that there is not a significant relationship between these two parameters.

1. **RESULT DISCUSSION**

8.1Logistic Regression:

The logistic regression model exhibits an ROC AUC of 0.80, indicating moderate predictive capability in discerning drinkers from non-drinkers. While offering valuable insights into drinking behaviour, accuracy of logistic regression is high compared to SVM. However, the choice between models depends on the dataset's complexity and the trade-off between interpretability and predictive power, with logistic regression excelling in linear relationships and capturing complex patterns. In logistic regression we are getting F1 score around 0.724 , recall is 0.73 and Precision is 0.71 (fig-15). which shows that the model is performs moderately well.

8.2. Support Vector Machine (SVM):

Both SVM and logistic regression achieve comparable ROC AUC scores, with SVM yielding 0.799, indicating effective performance. Both models can be considered successful in discerning drinkers from non-drinkers. F1 score of this model is 0.719, recall is 0.72 and precision is 0.716(FIG.17)

1. **CONCLUSION**

Based on the findings, the Logistic regression model emerges as the preferred option for predicting outcomes in the smoker-drinker dataset, given the current feature set and preprocessing steps. While the ROC AUC provides valuable insights into a model's class differentiation abilities, it's crucial to consider other factors like interpretability, computational efficiency, and overall model effectiveness when selecting a deployment model. While comparing F1 score, recall and precision logistic regression showing better performance.

A comparison of ROC AUC, F1 values indicates superior performance of the Logistic Regression over SVM model. This suggests that SVM may not adequately capture the intricate relationships within the dataset, particularly the linear version used here, potentially struggling to handle its inherent complexities.

1. **FUTURE WORKS**

In future iterations of the smoker-drinker dataset project, incorporating additional body signal data sensitive to alcohol consumption could refine outputs and enhance prediction accuracy. Exploring advanced machine learning models like neural networks or gradient boosting may further improve drinker classification. Additionally, refining the dataset through deeper feature engineering could uncover more intricate patterns. These approaches collectively hold promise for advancing the project's predictive capabilities and understanding of the complex relationships between alcohol consumption and physiological signals, opening avenues for more precise and impactful insights into drinking behaviour.

1. **SOCIAL AND ETHICAL IMPACT**

The use of advanced machine learning and extra physiological data to predict drinking behaviour raises concerns about privacy, stigma, and data bias. Ethical considerations include obtaining informed permission, reducing bias in model building, and resolving any inequities. Transparency, accountability, and fair access to interventions are critical for responsible implementation and minimising unexpected consequences. While predictive models have the potential for early intervention and public health advantages, careful consideration of social and ethical consequences is required to protect human rights, promote fairness, and reduce possible harm in society.

**Reference**

Domingos, P. (2012). A few useful things to know about machine learning. *Communications of the ACM*, *55*(10), 78. <https://doi.org/10.1145/2347736.2347755>

*StringIndexer — PySpark master documentation*. (n.d.). Spark.apache.org. Retrieved March 26, 2024, from <https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.feature.StringIndexer.html#:~:text=A%20label%20indexer%20that%20maps>

Batt, S., Grealis, T., Harmon, O., & Tomolonis, P. (2020). Learning Tableau: A data visualization tool. *The Journal of Economic Education*, *51*(3-4), 1–12. <https://doi.org/10.1080/00220485.2020.1804503>

Jr, D. W. H., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. In *Google Books*. John Wiley & Sons. <https://books.google.co.uk/books?hl=en&lr=&id=bRoxQBIZRd4C&oi=fnd&pg=PR13&dq=Hosmer>

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/bf00994018

*Logistic Regression with Pyspark: Customer Churn*. (n.d.). Www.youtube.com. Retrieved March 28, 2024, from <https://www.youtube.com/watch?v=r98gP73l9SU>

Drabas, T., & Lee, D. (2017). Learning PySpark. In *Google Books*. Packt Publishing Ltd. <https://books.google.co.uk/books?hl=en&lr=&id=HVQoDwAAQBAJ&oi=fnd&pg=PP1&dq=data+cleaning+in+machine+learning+pyspark+example&ots=tNGlMgIf9P&sig=jLtWIeyWbJPLx-QOe1uXjq9bW4c#v=onepage&q&f=false>

Marzban, C. (2004). The ROC Curve and the Area under It as Performance Measures. *Weather and Forecasting*, *19*(6), 1106–1114. https://doi.org/10.1175/825.1

‌

‌

**Appendix**

1. Screen shots

**PYSPARK**

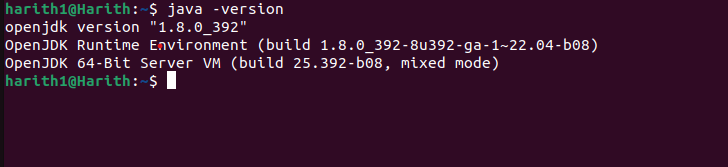


Fig-1 Exporting java to working terminal

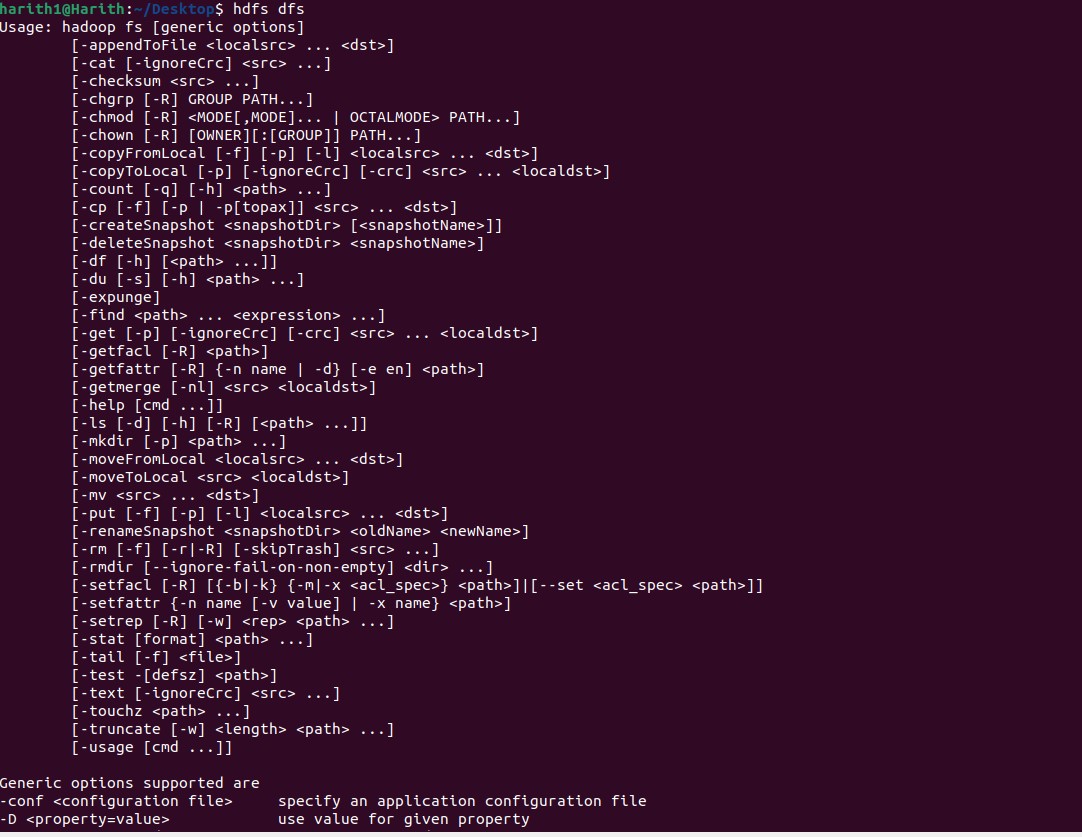


Fig-2.a Hadoop installation

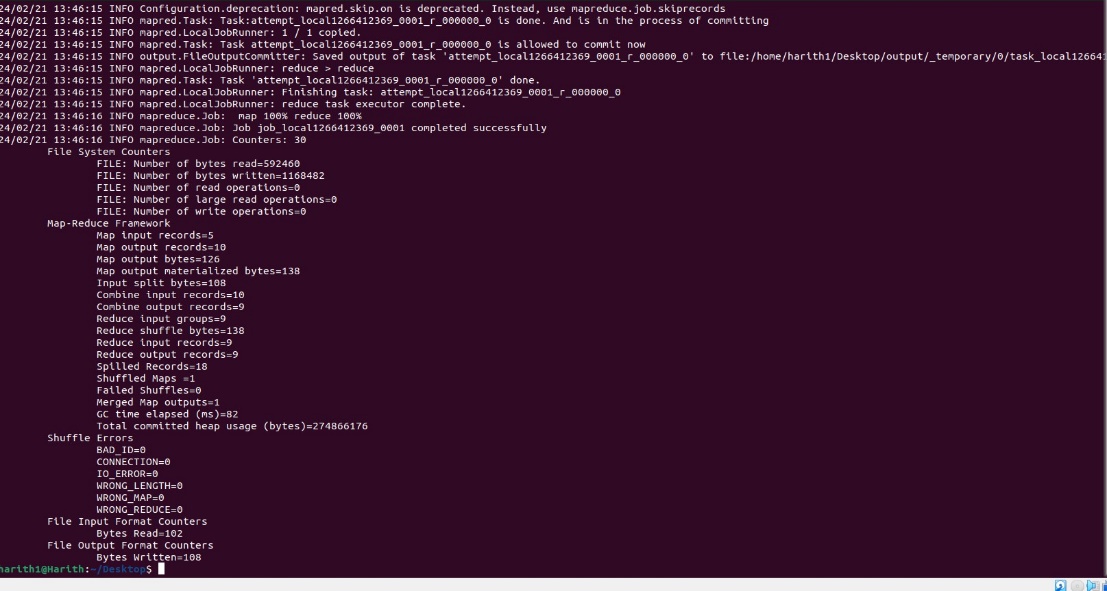
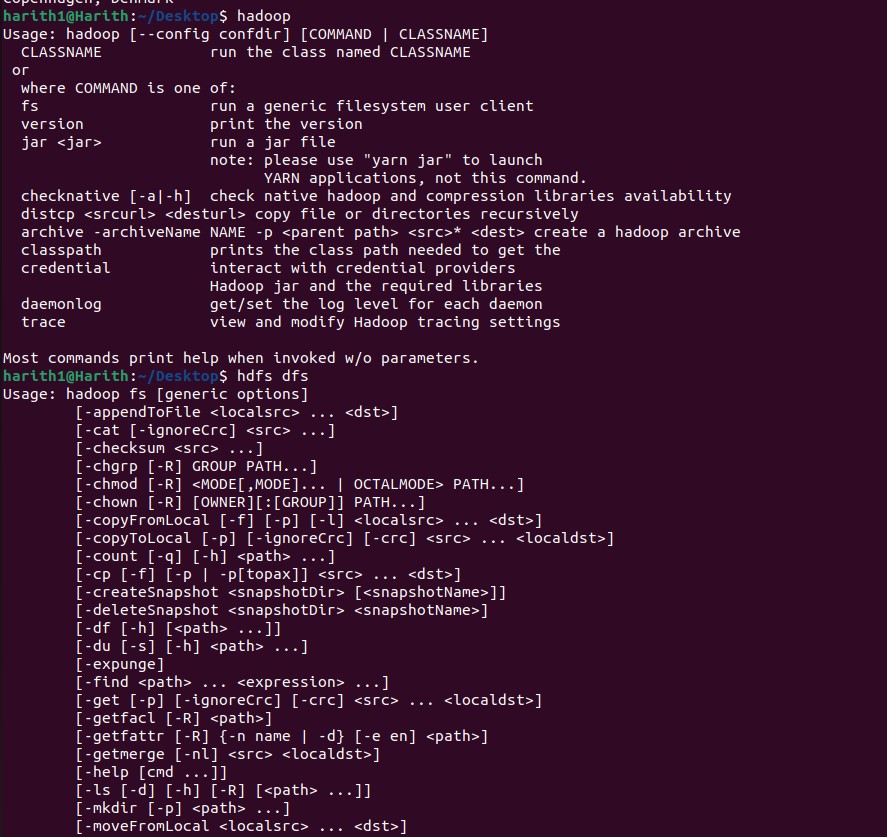


Fig-2.b Hadoop installation

Fig-2.c Hadoop installation

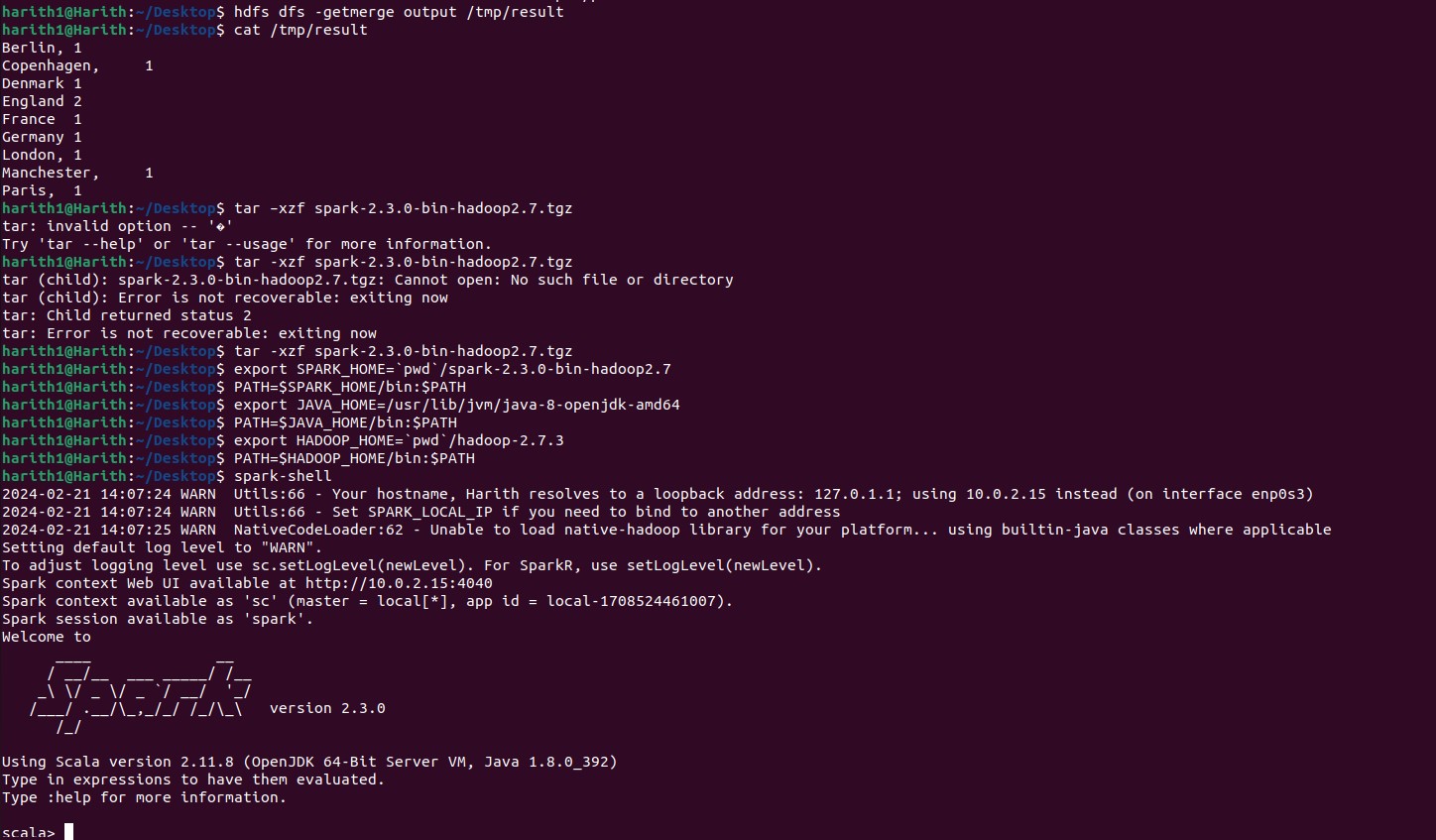


Fig-3.a Spark installation

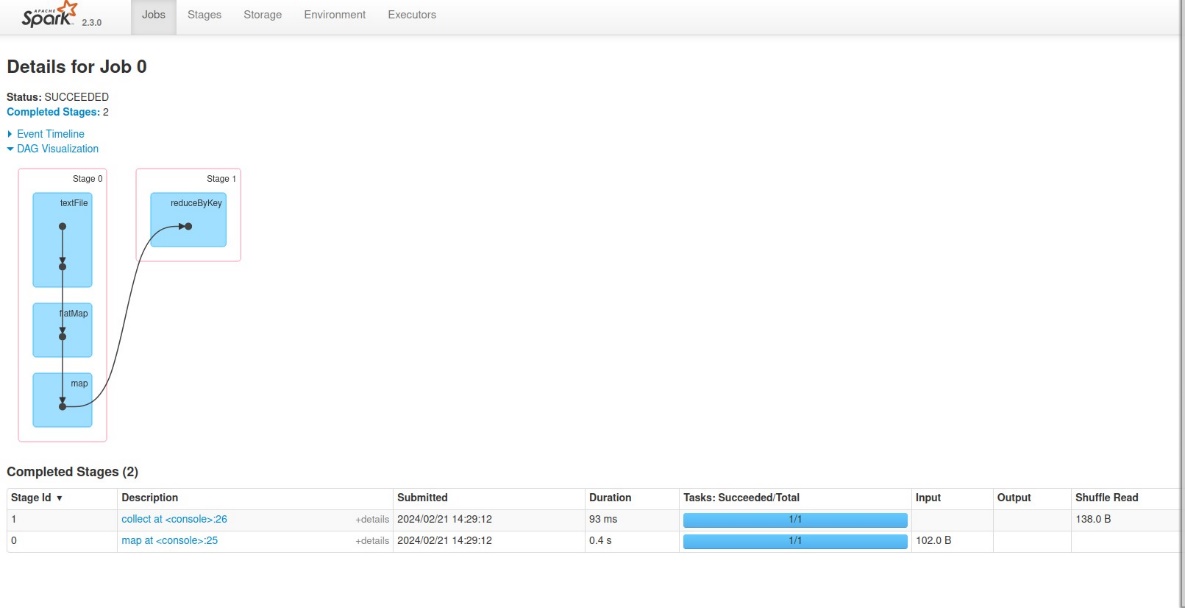


Fig-3.b Spark installation

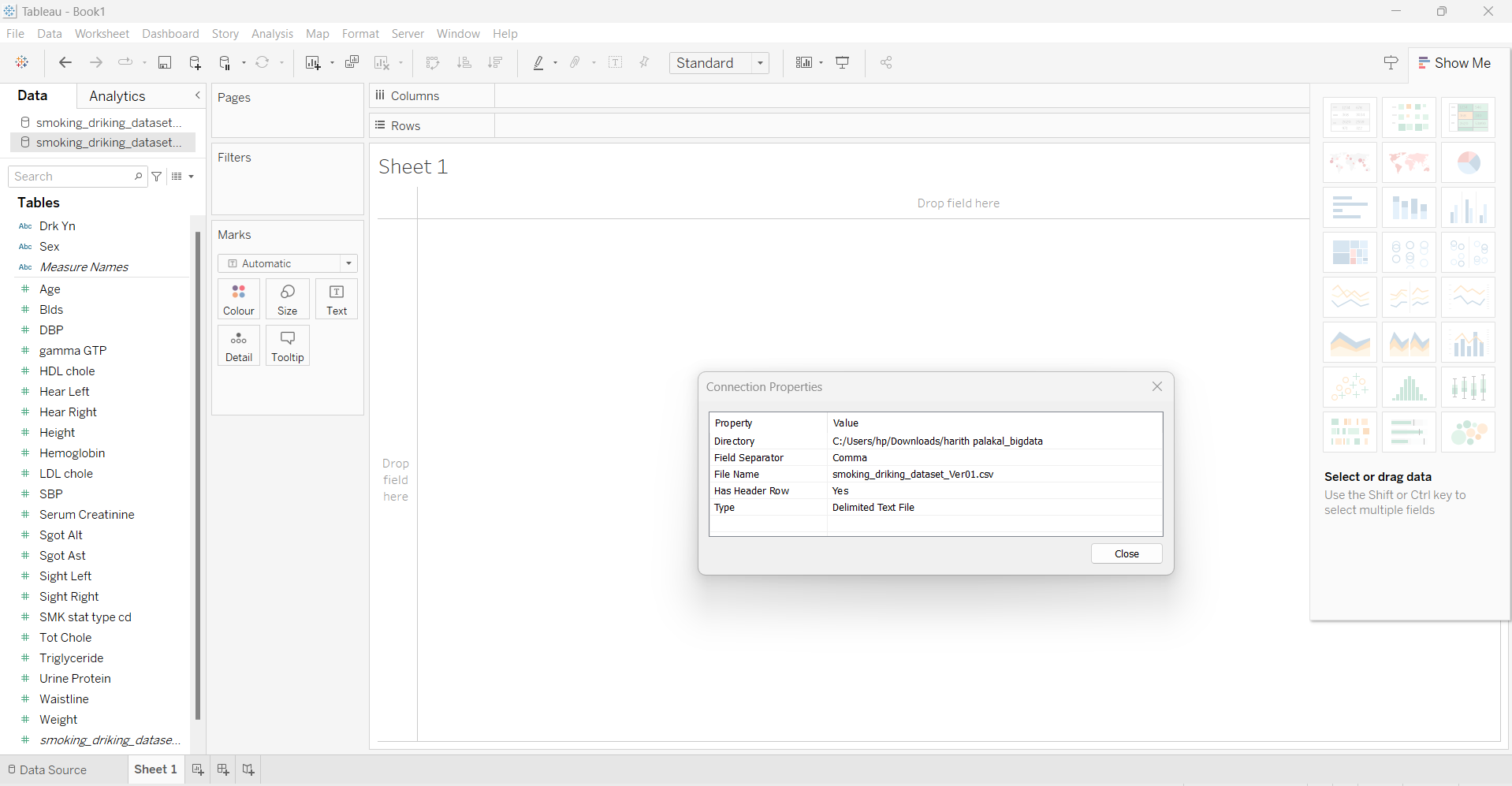


Fig.3.C Tableau installation

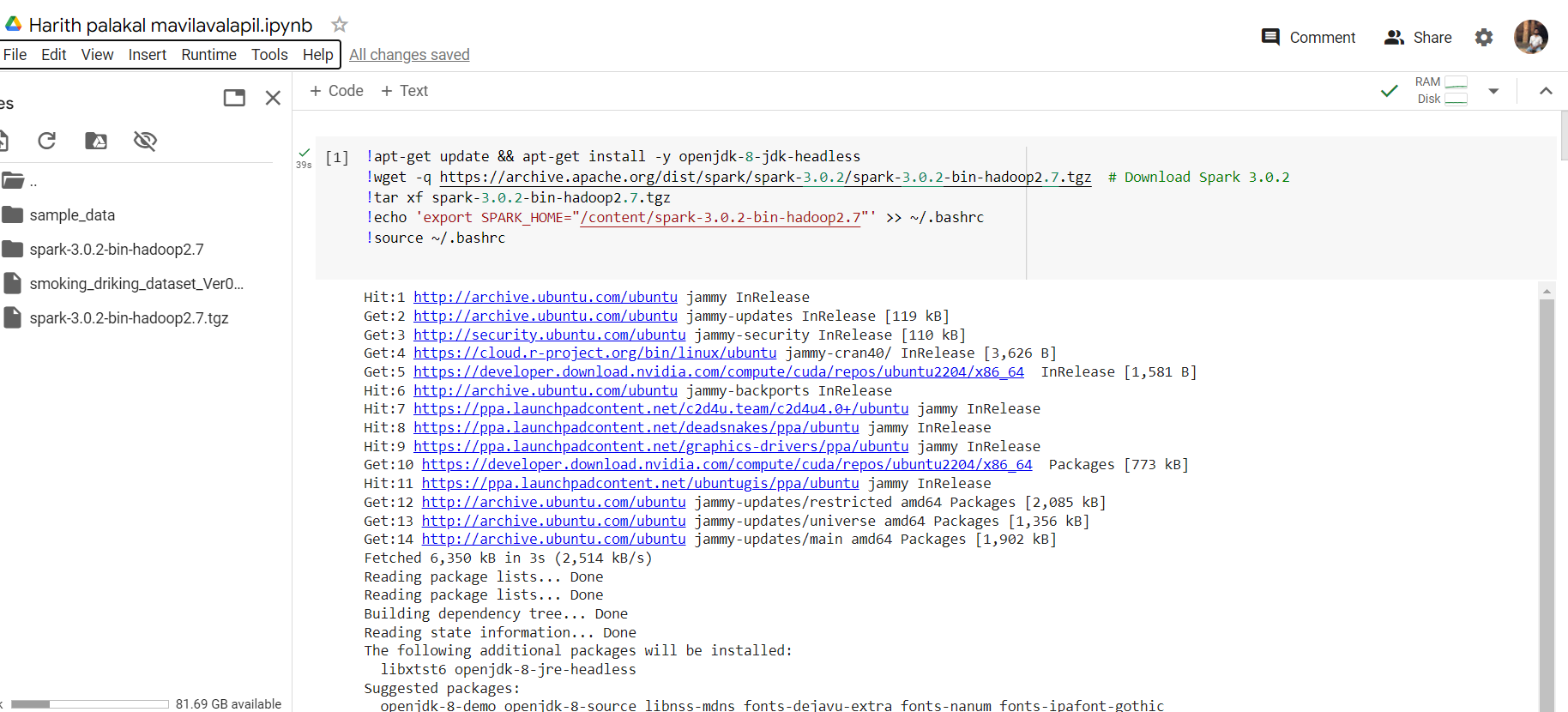


Fig-4 installing java ,spark in google colab

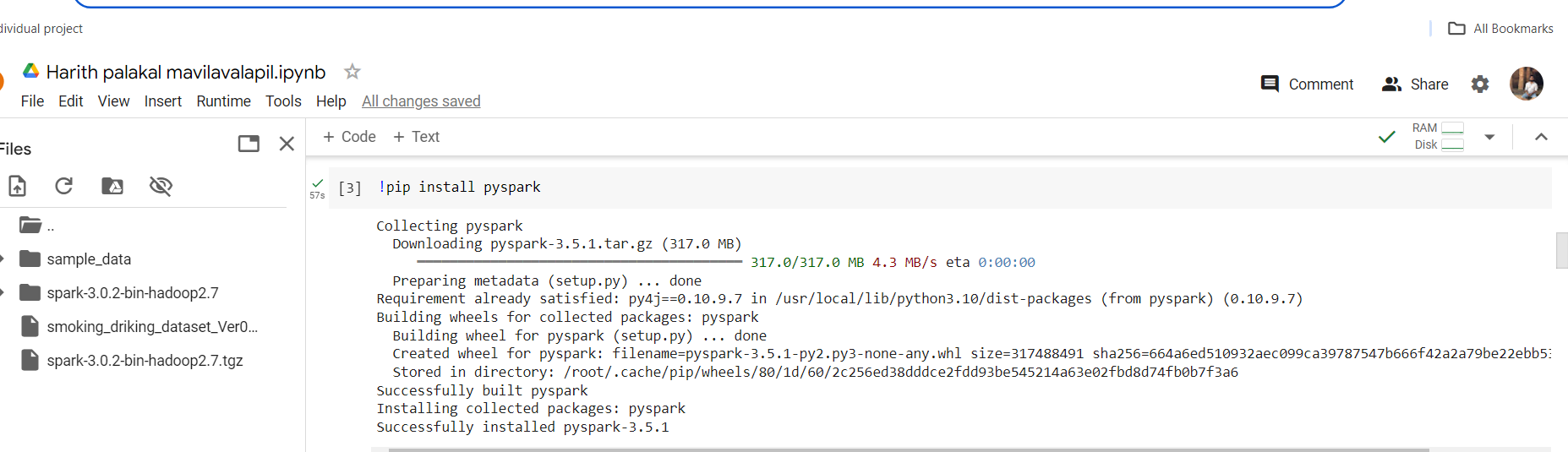


Fig-5 installing pyspark in google colab

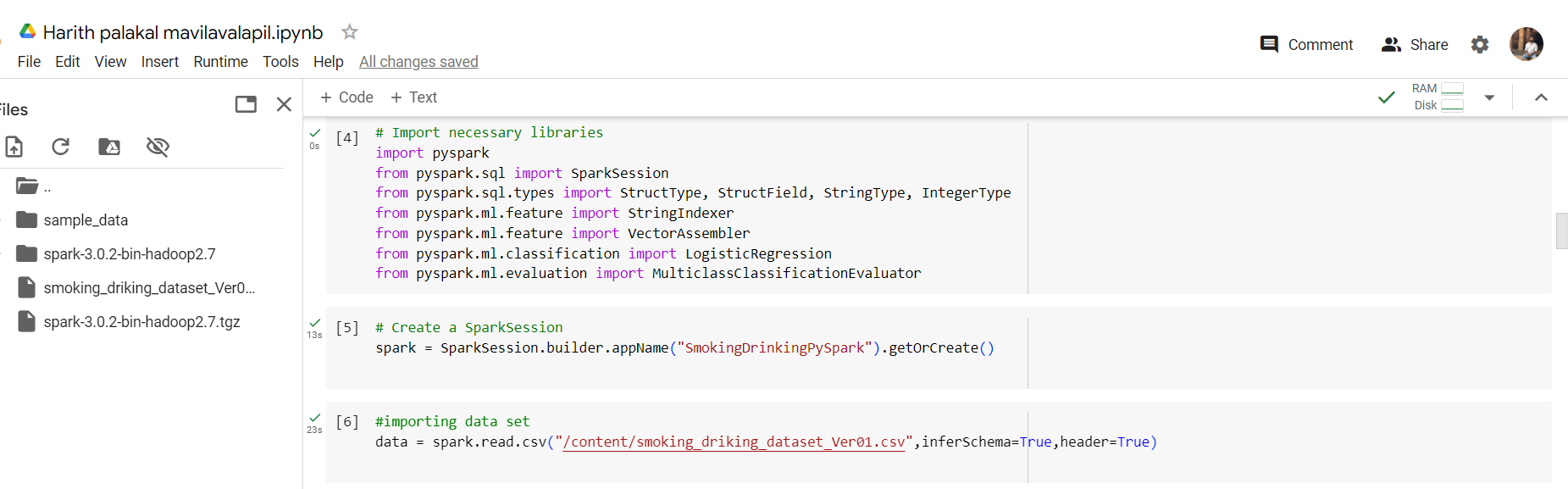


Fig-6 Importing libraries and creating spark session

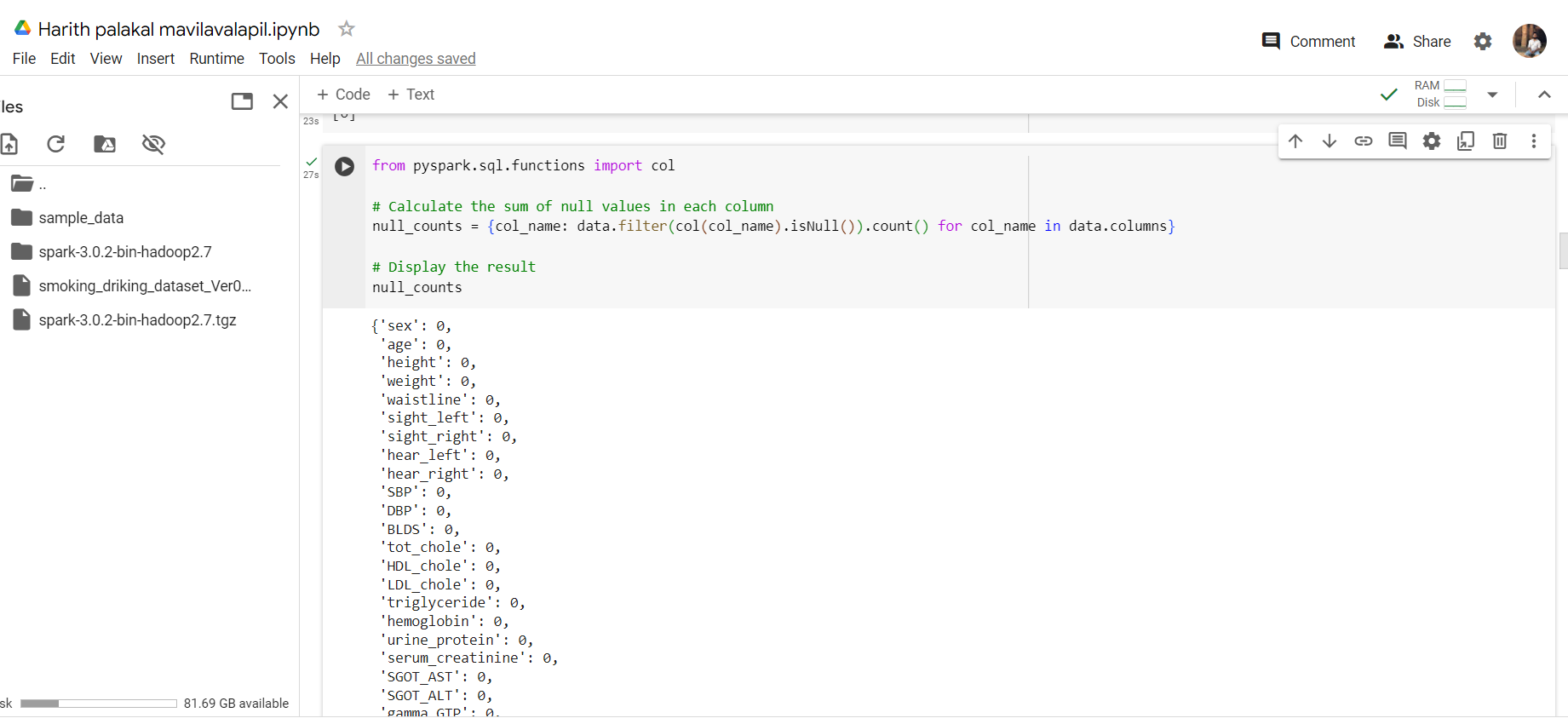


Fig-7 Calculate the null values in the column

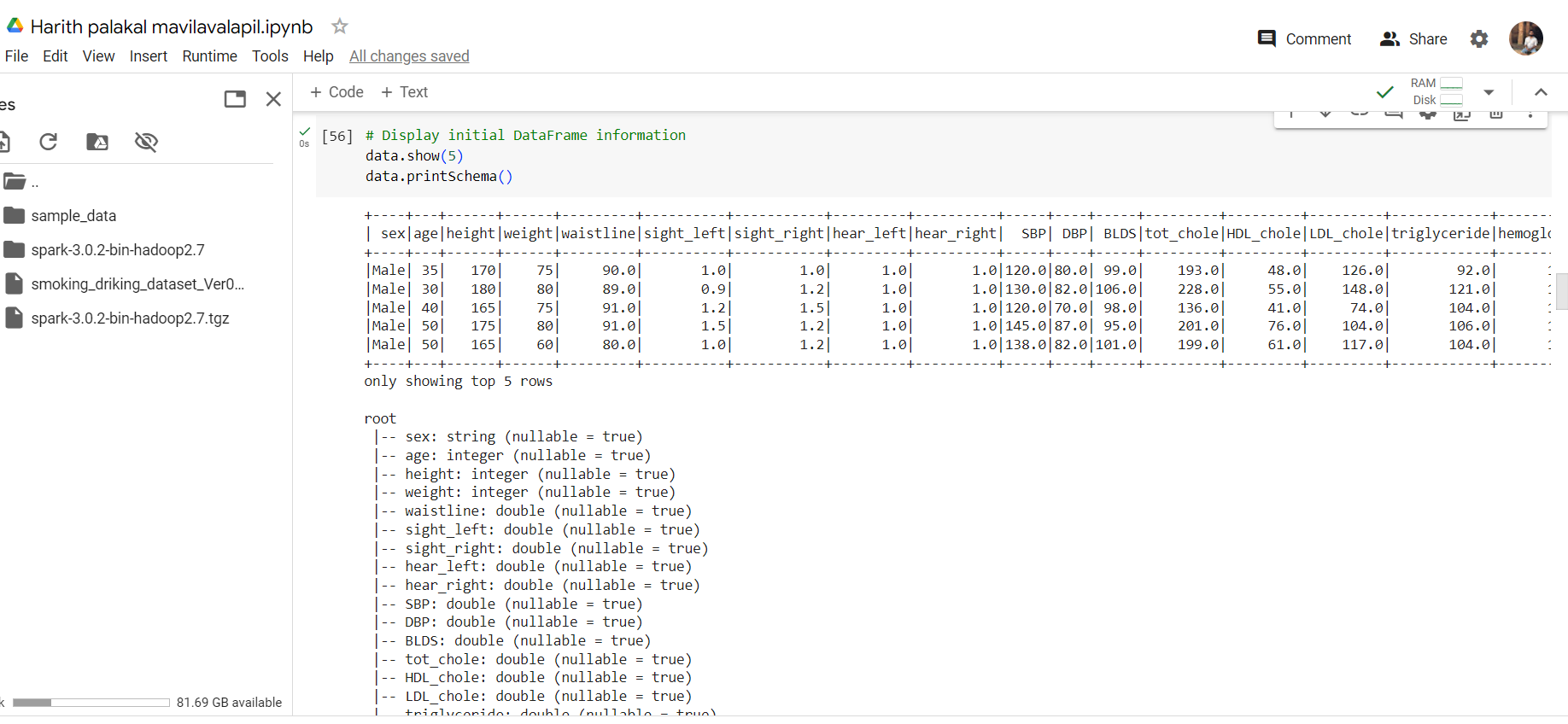


Fig-8 Displaying data and schema



Fig-9 Data encoding



Fig-10 Vector Assembler

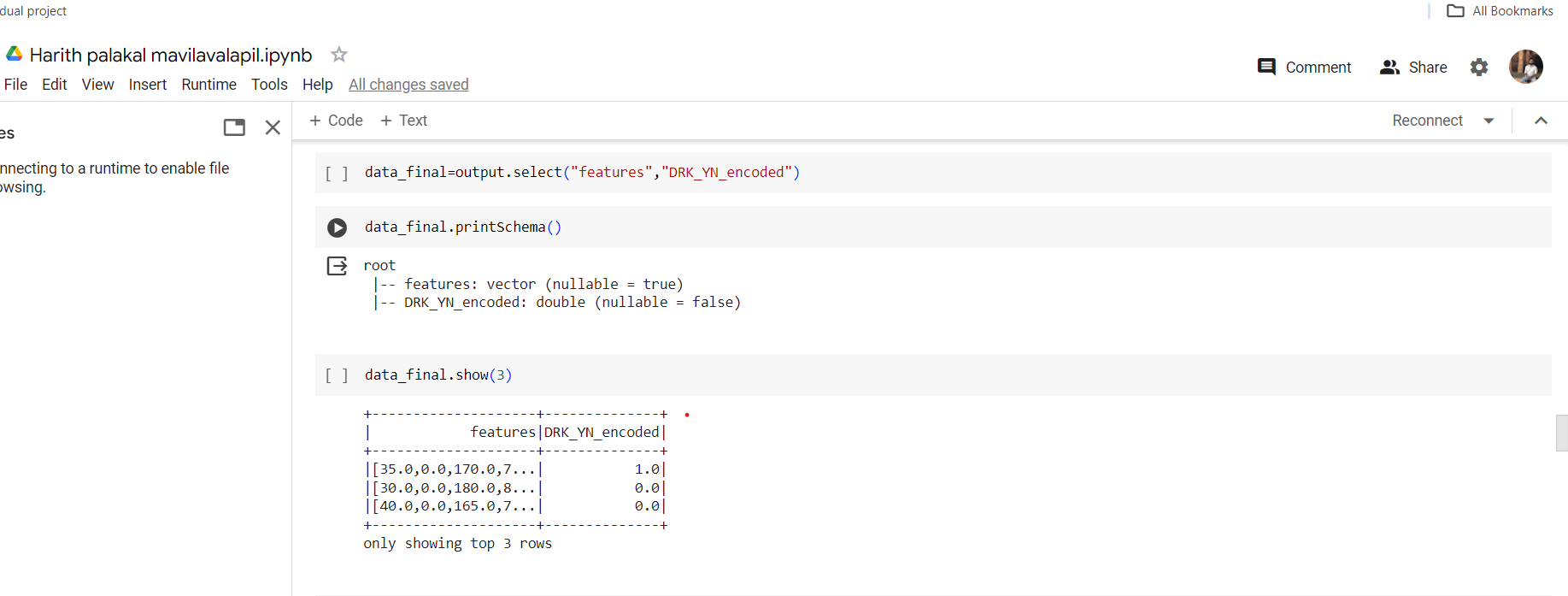


Fig-11 Creating new DataFrame

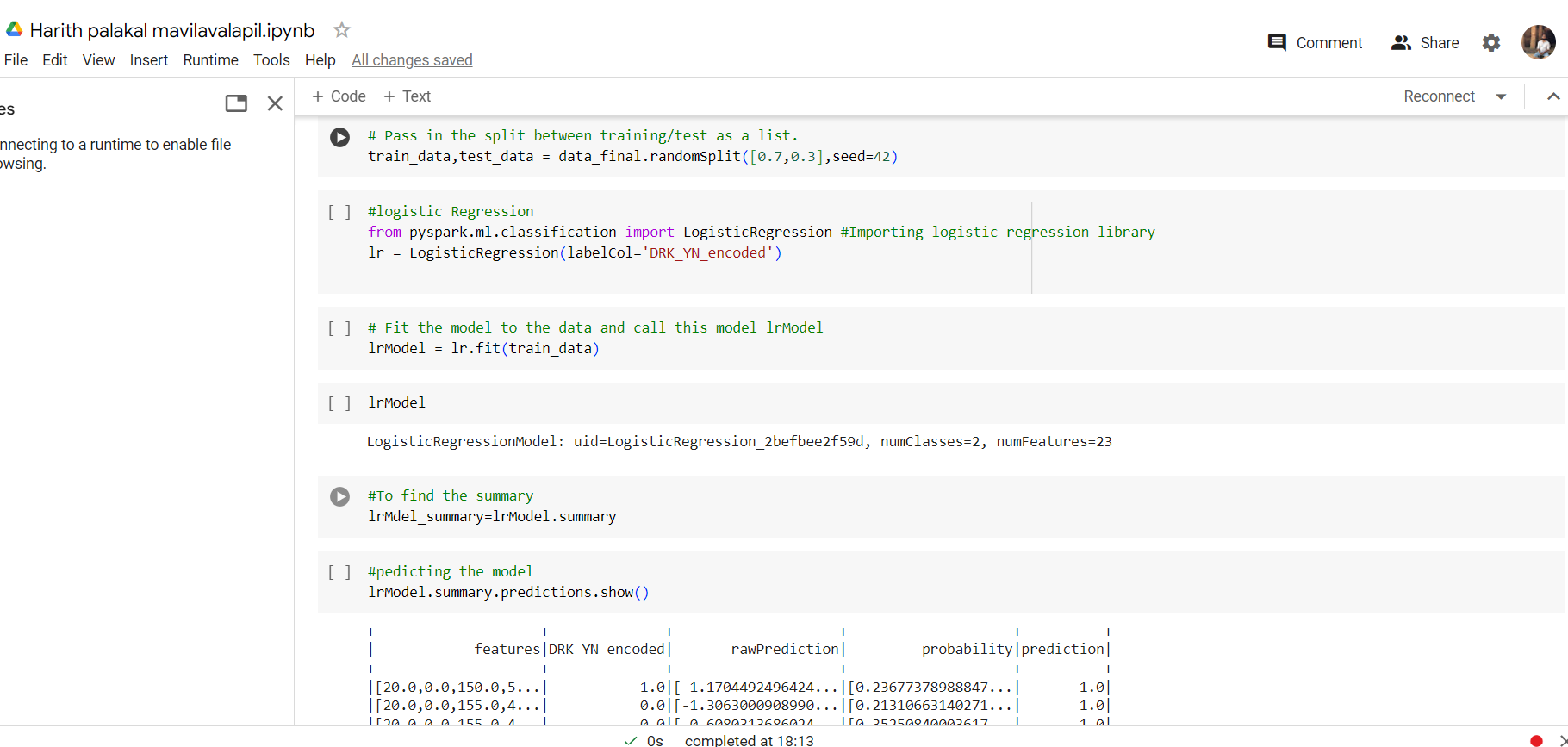


Fig-12 Splitting and Logistic Regression model creation



Fig-13 Accuracy of Logistic Regression model (ROC AUC)



Fig-14 F1 score, recall and precision of Logistic Regression model

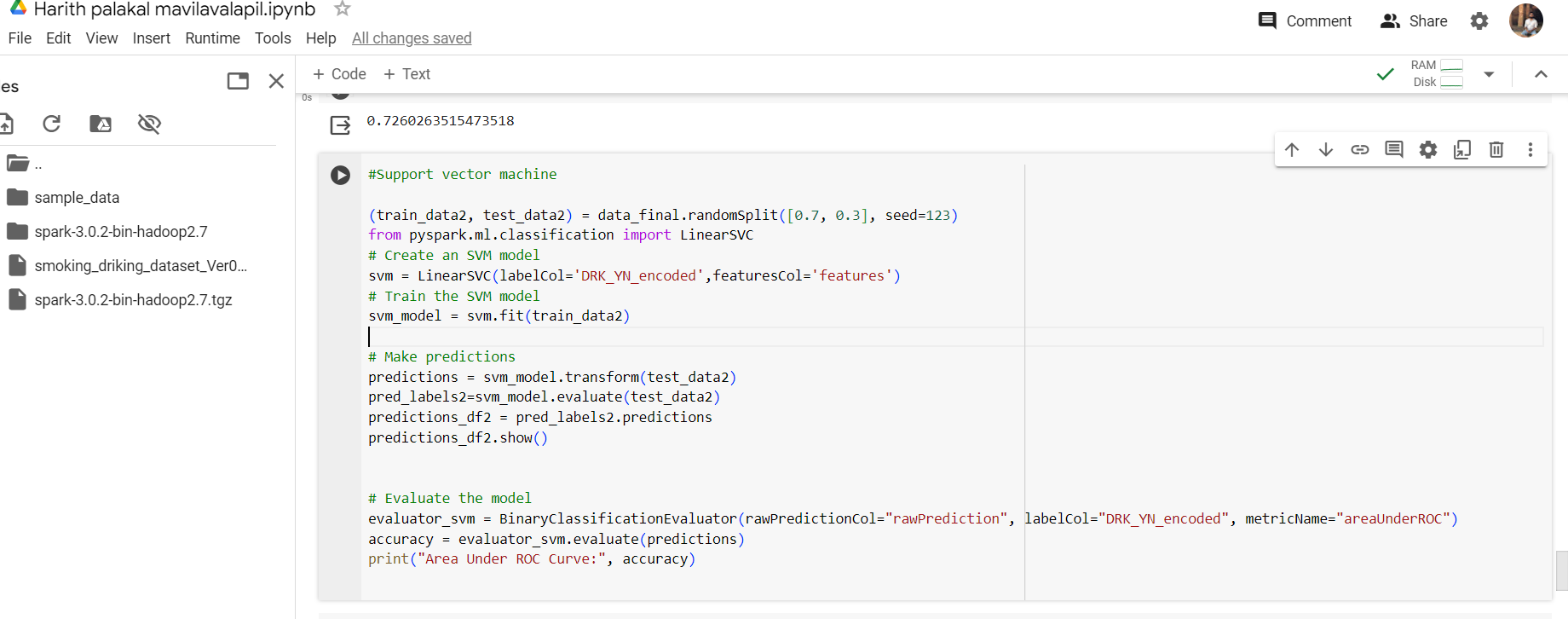


Fig-15 Creation of SVM model

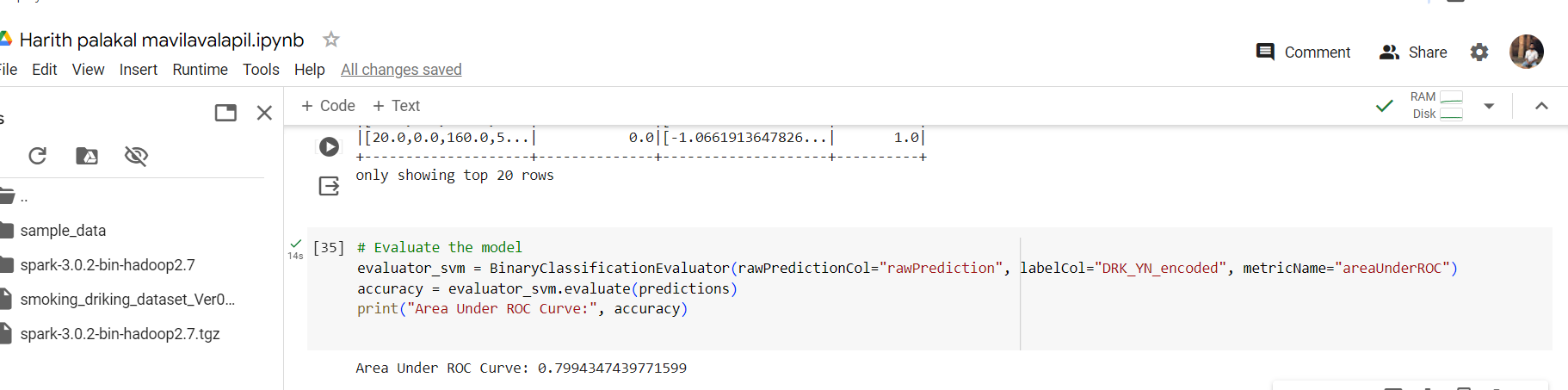


Fig-16Accuracy of SVM model



Fig-17 F1 score, recall and precision of SVM

**TABLEAU**

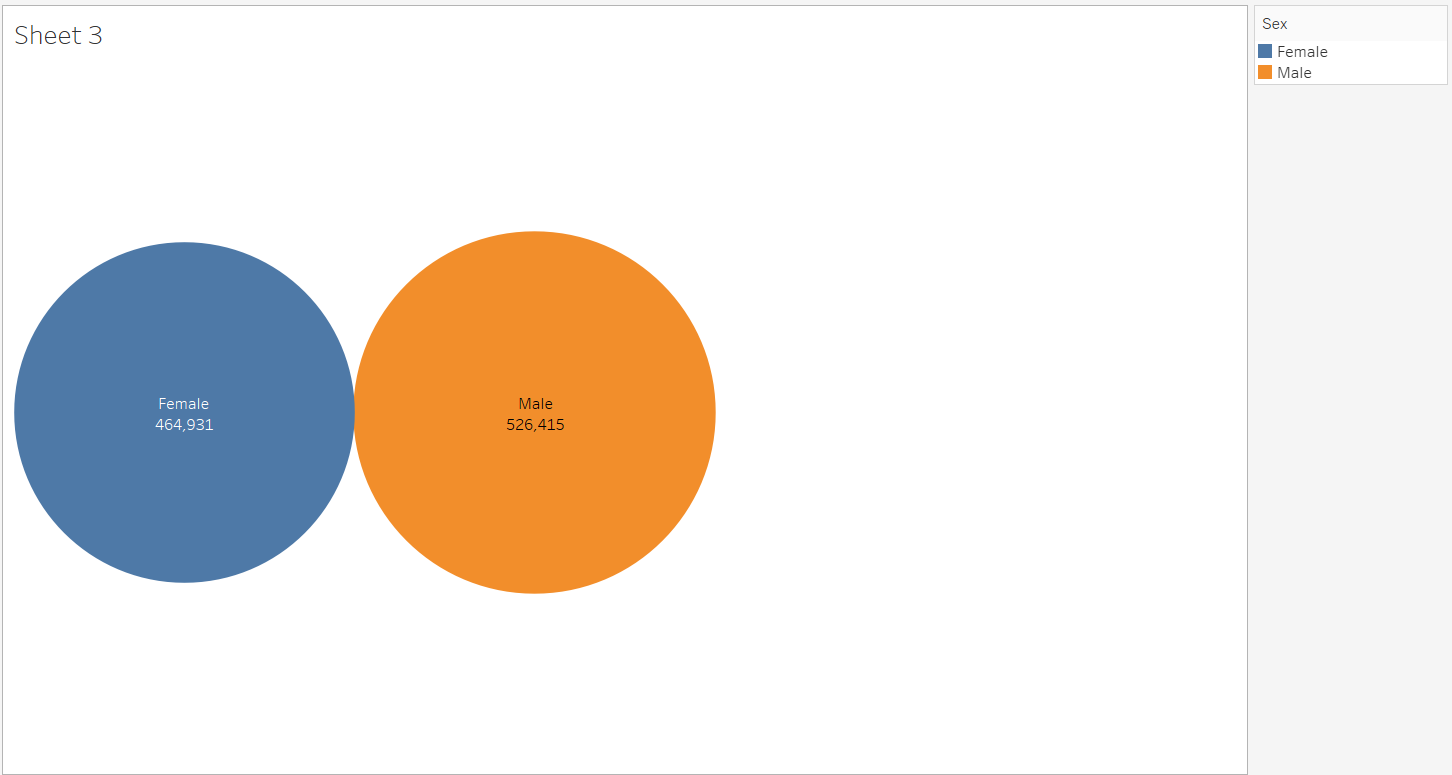


Fig-18 Total number of male and female

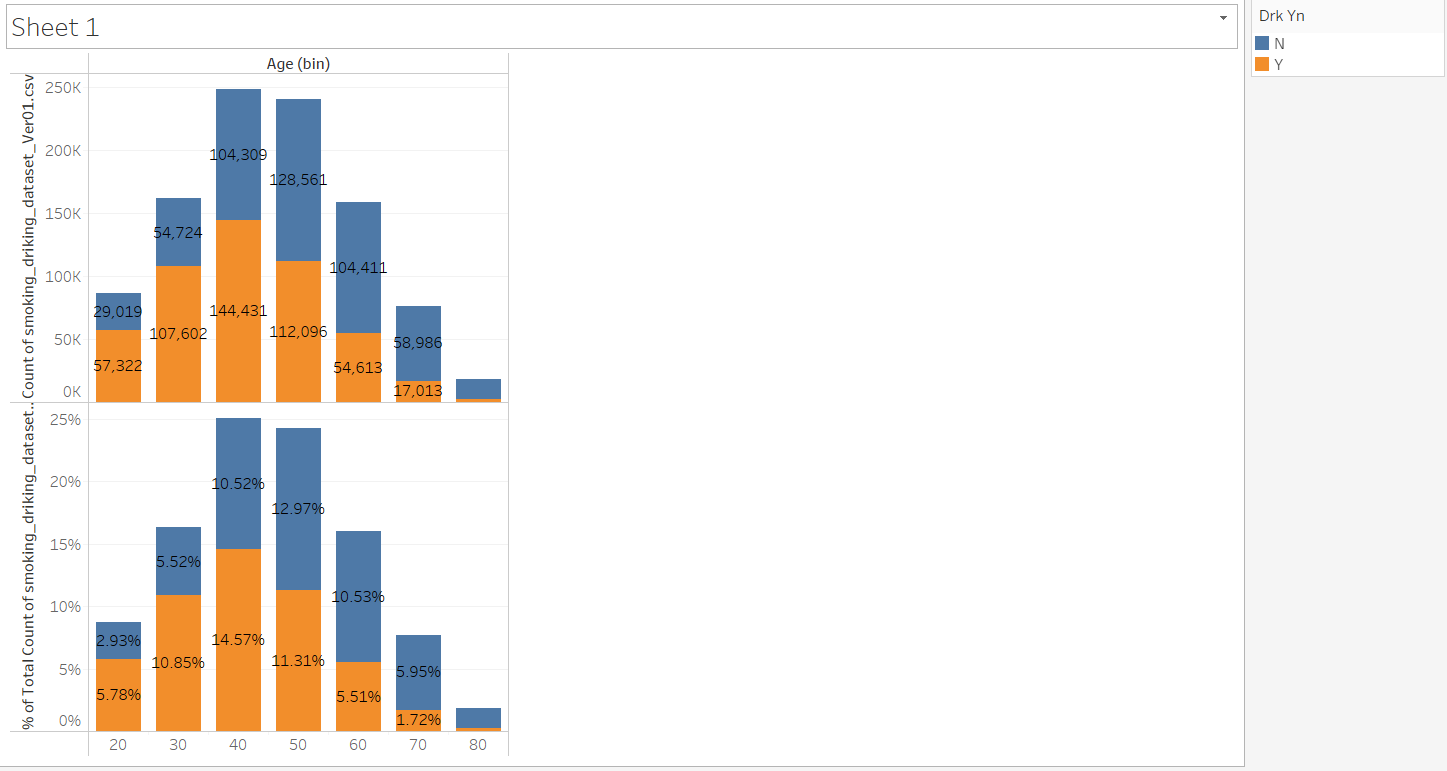


Fig-19 Age vs Drnk Yes/No

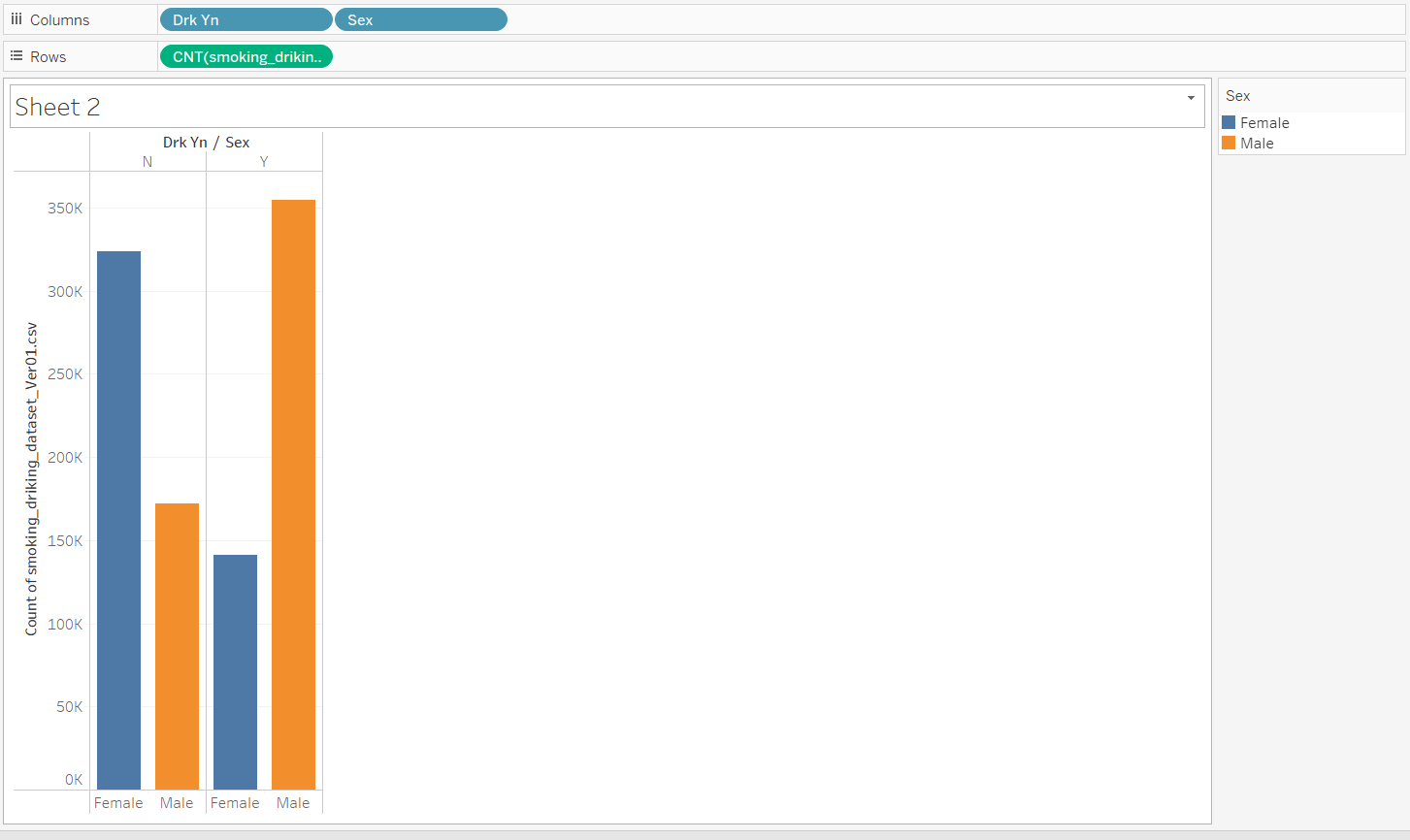


Fig-20 Sex vs Drnk Y/N

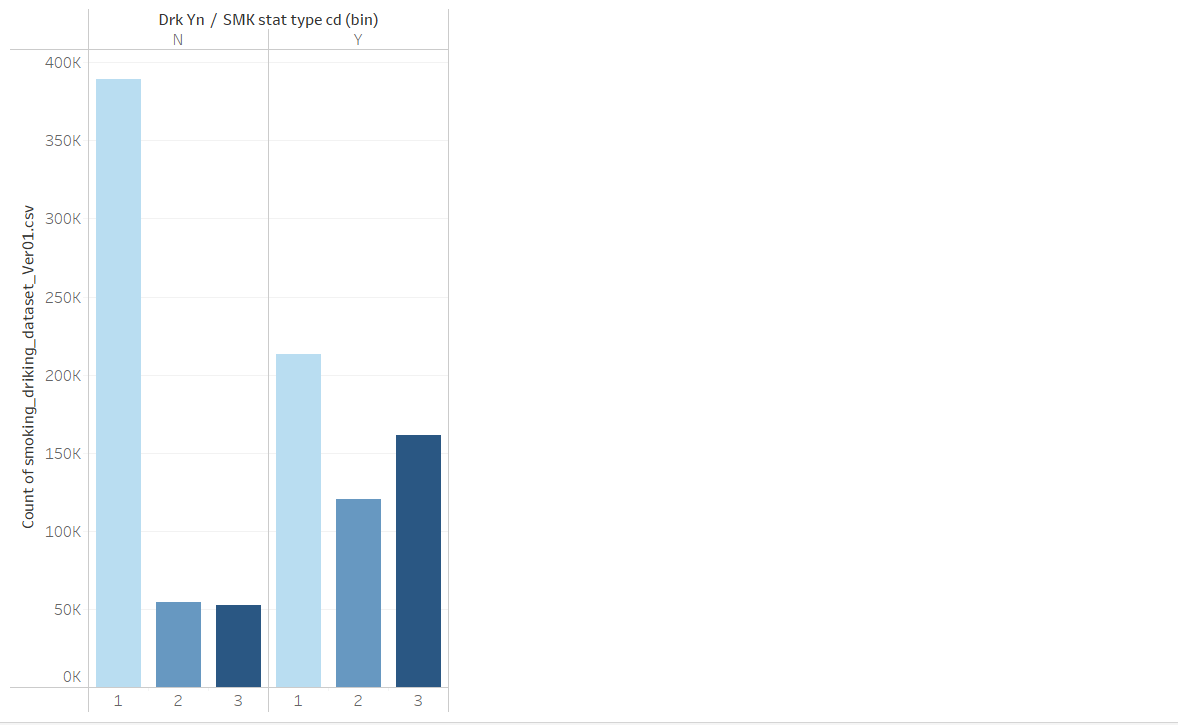


Fig-21 Drinking Vs Smoking

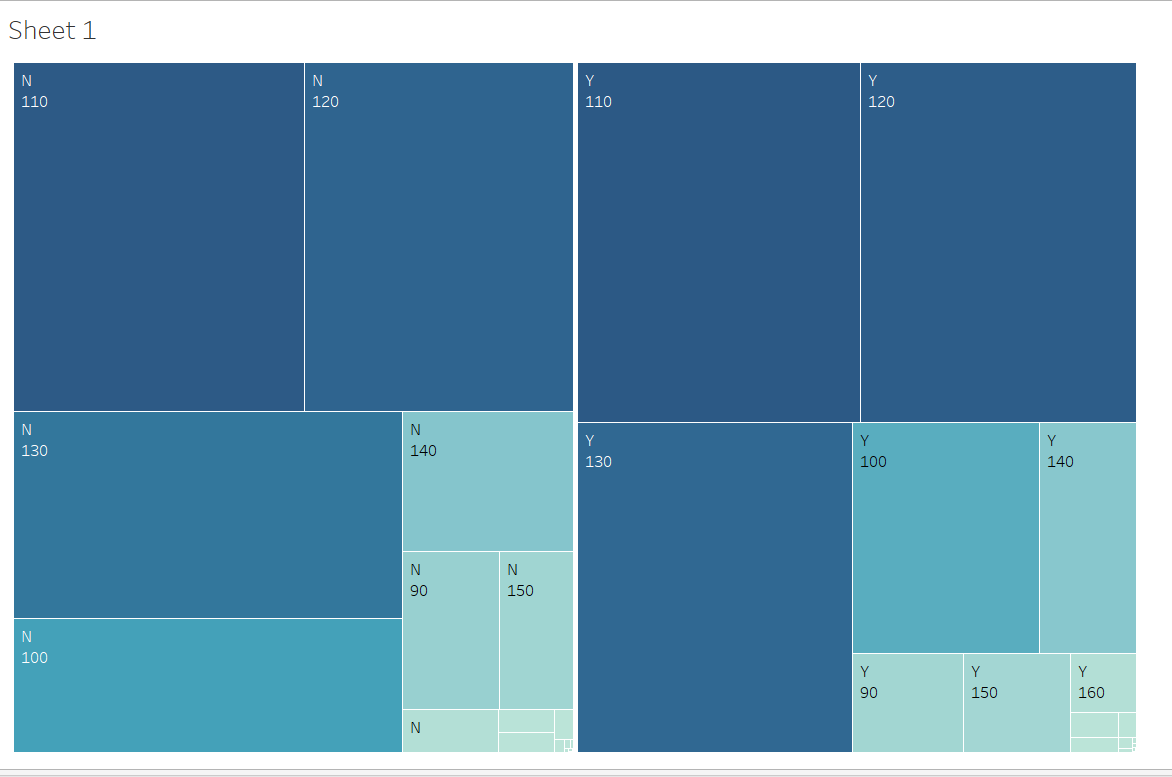
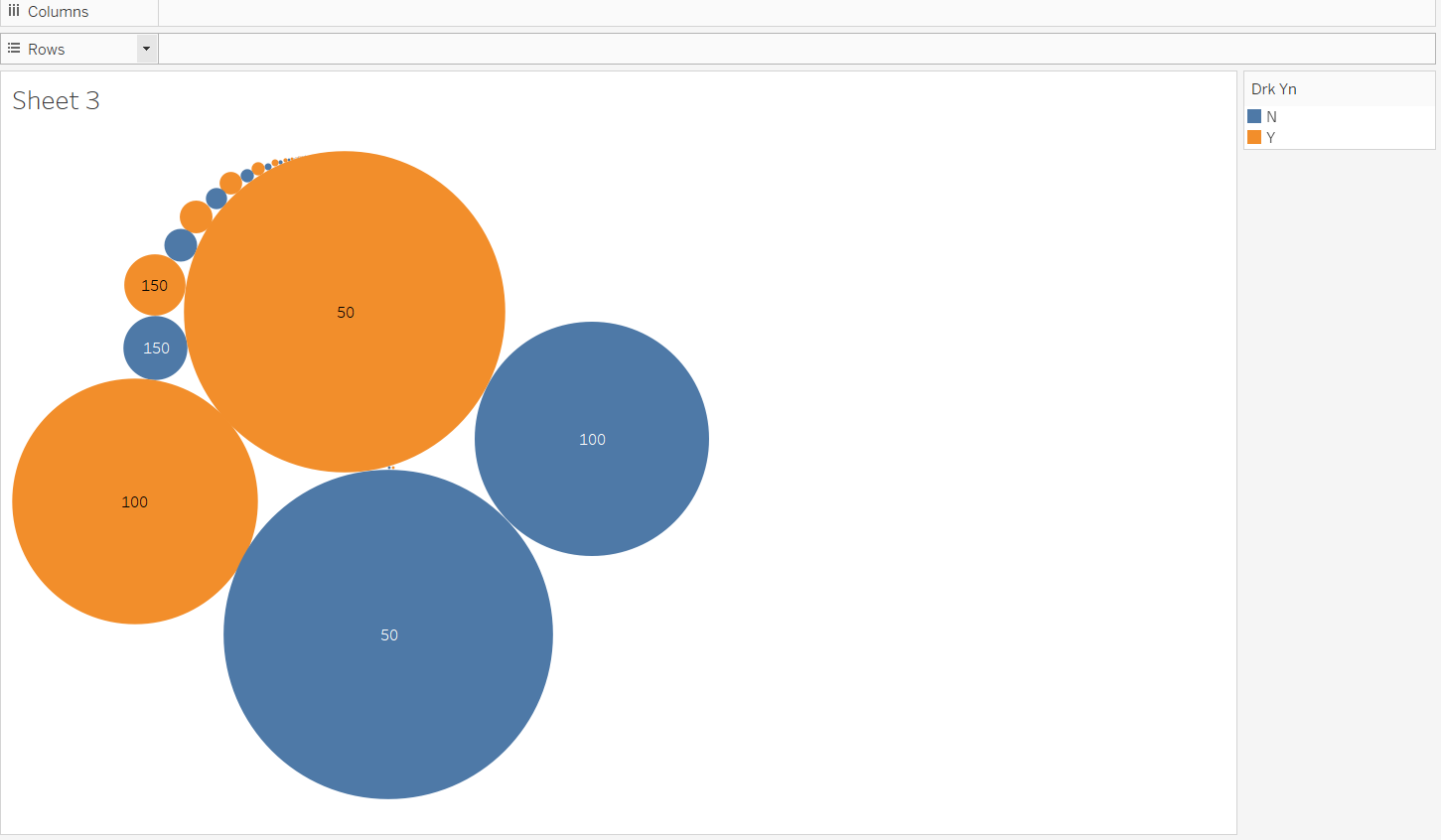


Fig-22 SBP vs Drinker Y/N



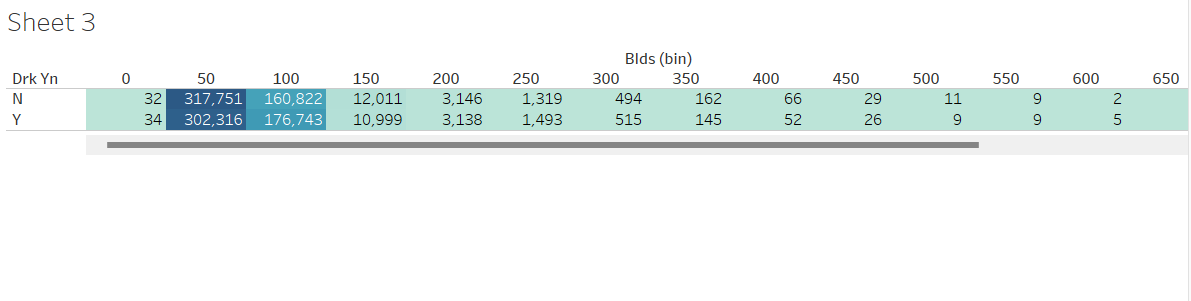


Fig-23 BLDS or FSG(fasting blood glucose) vs Drinker Y/N