

RESOURCE OPTIMIZATION BY PREDICTING CROP PRODUCTION

Submission by:
Nithyashree Arunachalam

Under the Guidance of
Mr. Farman Khalid

INDEX

TOPIC	PAGE NUMBER
Abstract	2
Objective	3
Introduction	4
Tobacco Industry in India	7
Resource Allocation in FMCG Companies	8
Data Analytics	9
Predictive Analytics	10
Forecasting	11
Time Series Forecasting Exponential Smoothing	12
Holt's Method	13
Language and Tools	14
Database of Farms	15
Forecasting Data	19
Creating Visualisation	25
Conclusion	27
Summary Recommendation	29
References	30

ABSTRACT

FMCG COMPANIES

The main feature of these goods is that:

- They have a short shelf life because of high consumer demand or because they are perishable.
- These goods are purchased and consumed frequently and rapidly.
- They are priced low and are sold in large quantities.
- They have a high turnover when they're on the shelf at the store.

A company that produces these goods are called FMCG companies. FMCG companies must sell as many units as they can as quickly and as consistently. These goods have a small profit margin along with a short shelf life in many cases. This requires clever marketing and high product quality.

Resource allocation is the process of allocating the managing assets in a way that aligns with the institution's strategic goals. Big companies like GPI are usually dealing with multiple projects. Effective allocation of resources helps project managers to assign resources for the project and manage them effectively. This way the projects can be handled in a much-organised way.

Right from manufacturing to delivery to marketing and sales, an FMCG company faces several challenges in the decision-making process and the only resource that can be relied upon is data. The company collects data in the form of POS data, customer demographic data, market research data, promotional campaign data, finance data, Twitter data, Facebook data, weather data etc. This data is usually unstructured, scattered and available in different formats. This data can be used for analysis which can be used for resource allocation, efficient marketing, boosting sales and solving a variety of other business issues using **Data Science and Machine Learning**.

The data of the amount of crop produced by farms in this case Tobacco crop is. Using machine learning algorithms, predictive analysis can be done to predict crop production for the coming year. This prediction will give us the idea as to how much produce each and every farm will approximately produce. Along with this, the other variables contributing to the cost of cultivation are also predicted. Visualization of this data and further analysing it will help us to draw a conclusion between production and total cost.

OBJECTIVE

Aim: Resource Optimization by predicting crop production

The objective of this project is to the concept of data science and machine learning to predict crop production which will help in allocating resources to these farms. In this project, we will focus on Tobacco production which is one of the most important cash crops in India and is used in a variety of products like cigars, cigarettes, and pipes.

Datasets containing the data of the crop produced and the resources used for agricultural activities by farms is taken for 7 years. With the help of data science and machine learning algorithms, predictive analysis can be done to predict crop production for the coming year. This prediction will give us an idea as to how much tobacco every farm will approximately produce. Along with this, the other variables contributing to the cost of cultivation are also predicted. Visualization of this data and further analysing it will help us to conclude the relationship between production and total cost for these farms. Resources are optimised when the amount of money or the total cost gone into cultivation is less with a good production rate. It would be a desirable outcome if the total cost is decreasing with increasing or decreasing production.

INTRODUCTION

About Godfrey Phillips India Limited

Godfrey Phillips India Ltd. (GPI) is a tobacco manufacturer headquartered in India. The firm was originally established in London in 1844. GPI was one of the first UK companies to mass-produce cigarettes, apart from being one of the founding companies of Imperial Tobacco along with John Player & Sons. GPI manufactures and sells cigarettes, smoking tobacco and cigars, apart from having a non-tobacco line of products released in 2009 that include confectionery.

GPI is the flagship company of Modi Enterprises, one of the largest cigarette manufacturers. The company has an annual turnover of approximately 7200 crores, according to 2018-19 estimate. The company also has business interests in pan masala, chewing and confectionery products.

Godfrey Phillips India recently launched Pan Vilas pan masala for the Indian market. Additionally, the company also manufactures and distributes Marlboro in India under a license agreement with Philip Morris.

Godfrey Phillips India's operations primarily span the country with prominence in the northern and the western part of the country. GPI has three manufacturing facilities spread across the country - at Rabale (Mumbai) and two units in Ghaziabad (near Delhi). They also have a state-of-the-art R&D centre in Mumbai and a food R&D in Ghaziabad, and a tobacco-buying unit in Guntur (Andhra Pradesh).

Godfrey Phillips India has an expansive International Business Division, that caters to all of GPI's product categories in International Markets. IBD has business associations with various players in the international tobacco industry to export its own cigarette brands, cut & blended tobacco, tobacco leaf and providing technical services and contract manufacturing. Many countries from the Middle East, Africa, Asia, Europe, Australia and Latin America are part of its portfolio.

Implementation of Data Science and ML at GPI

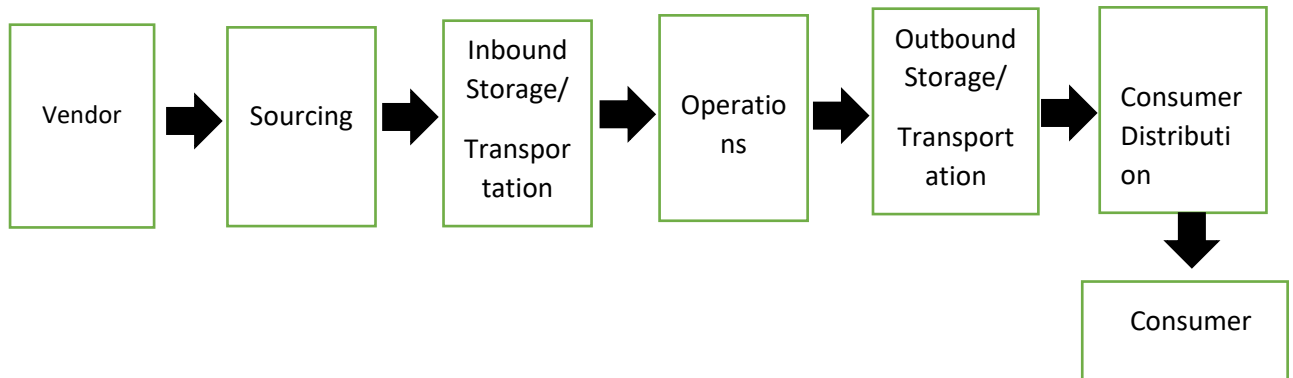
Data science and machine learning algorithms are being implemented in different ways in the organisation

1. Machine Learning for accurate event prediction
2. Sales and stock forecasting
3. Optimized Route Planning
4. Trade spend Analysis in FMCG

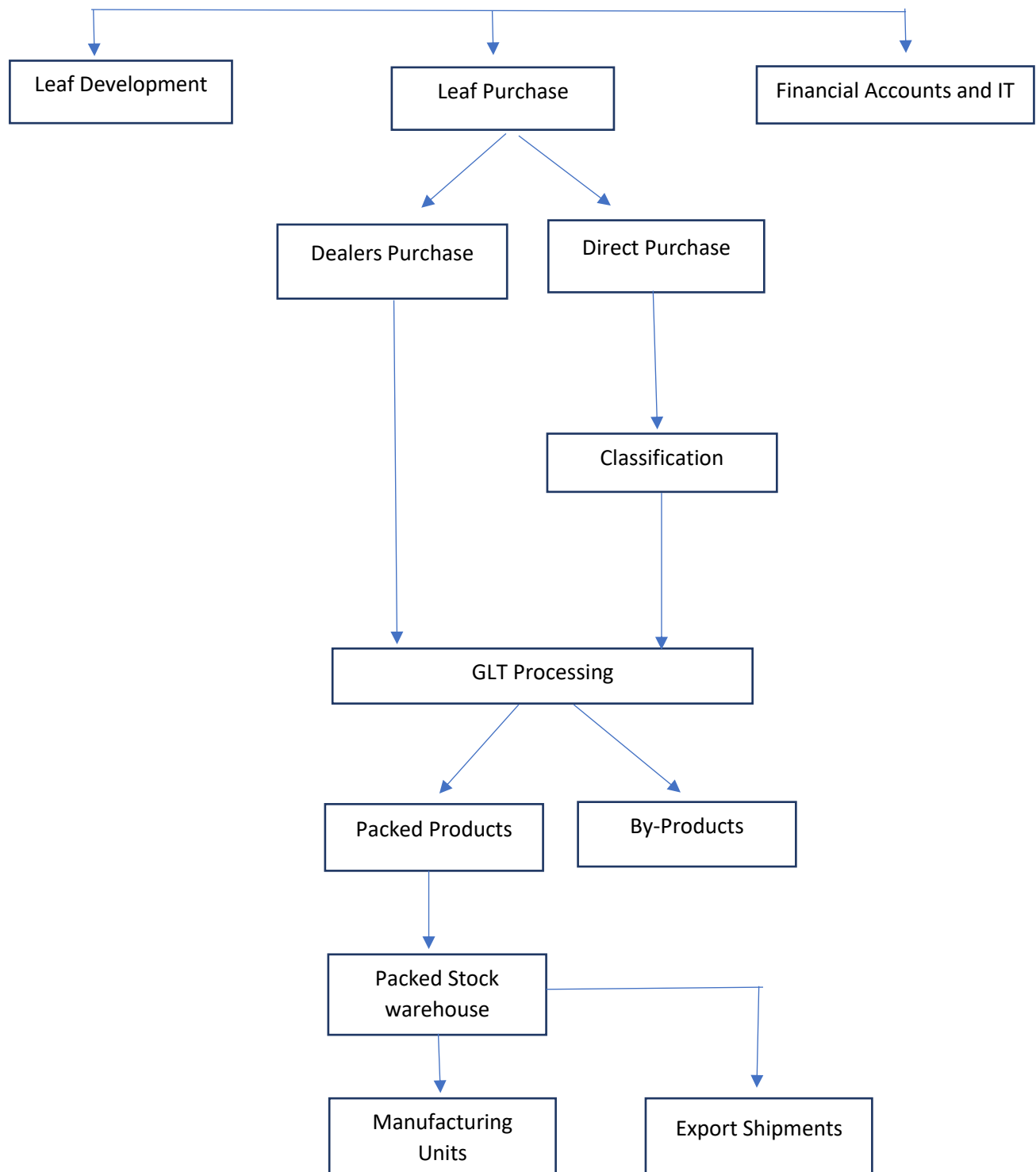
Supply chain management

Godrej Philips India Ltd has its own “Leaf Division” in the tobacco growing area of Guntur in Southern India, which gives the company access to the best quality of tobaccos.

A typical supply chain in an FMCG company is:



Leaf Operation Flowchart



**GLT->Green Leaf Threshing

TOBACCO INDUSTRY IN INDIA

Tobacco was introduced by the Portuguese in the 17th century in India. Tobacco is a plant grown for its leaves, which are dried and fermented before being put in tobacco products. The crop is often appreciated worldwide for its rich, full-bodied flavour and smoothness. It has now become a global commodity with India being the second largest producer and exporter in the world. It is needless to say that tobacco is an essential cash crop in the country providing livelihood to a number of farmers.

Tobacco Products in the market

People can smoke, chew, or sniff tobacco. Smoked tobacco products include cigarettes, cigars, bidis, and kreteks. Some people also smoke loose tobacco in a pipe or hookah (water pipe). Chewed tobacco products include chewing tobacco, snuff, dip, and snus; snuff can also be sniffed.

There are Ninety-three varieties of tobacco including FCV (29), Burley (3), Natu (5), Lanka (2), Chewing (17), Bidi (15), Cheroot (3), Cigar (4), Hookah & chewing (15) types have been released for farming community.

In India, tobacco is harvested mainly in two forms - cigarette tobacco and non-cigarette tobacco. During 2017-2018, non-cigarette tobacco alone had a 69% market share. Based on consumption, khaini constituted ~11%, and beedi and cigarette had a market share of 8%. Usage of smoking mediums like hookah, hookli, chhutta, dhumti and chillum, along with edible tobacco like mawa, snuff, gutka, and pan masala have led to the growth of this market.

Tobacco Board

The Tobacco Board of India is a facilitator for tobacco growers, traders and exporters. The main objective of the board is to estimate demand and regulate the production of FCV (**Flue Cured Virginia**) tobacco to match the demand to ensure a fair price for the produce. It extends help to the tobacco farmers in securing crop loans, quality seeds, fertilisers and other crucial inputs. They also provide guidance to the framers on GAP (Good Agricultural Practices) to produce quality tobaccos to meet the current international standards.

RESOURCE ALLOCATION IN FMCG COMPANIES

Resource allocation is the process of allocating the managing assets in a way that aligns with the institution's strategic goals.

Importance of effective resource allocation

Big companies like GPI are usually dealing with multiple projects. Effective allocation of resources helps project managers to assign resources for the project and manage them effectively. This way the projects can be handled in a much-organised way.

1. **Save money:** Effective resource allocation reduces the wastage of money. It helps monitor the performance of team members in a project. Hence makes it easier to assign resources to those stages in the process and to those teams who need it while reducing it from areas where it is not being used to its maximum potential.
2. **Boost productivity:** It is the primary reason to choose resource allocation. Resource allocation helps you to know who is overloaded and who is free at that instant.
 1. **Improve time management:** It is important to know how long it takes the resources to complete the projects or tasks, to run a project efficiently. Proper allocation of resources can set the actual estimate hours to complete the tasks.
 2. **Strategic planning:** Resource allocation plays a vital role in helping to achieve the organisation's goals. Proper allocation of resources can help to achieve and fulfil project needs.

Challenges due to inefficient resource allocation:

1. Resources are assigned inconsistently: Allocating resources without properly studying the incoming demand, setting priorities and considering the company's goals can lead to wastage of these resources.
2. Shifting resources in response to unexpected problems: unexpected issues and risks might crop up at any given point in time. That doesn't mean shifting resources here and there sporadically should be the solution. A good grasp on demand and requests is required along with the creation of a planned ideally under a robust project management methodology.
3. Resource use is not optimized: Sometimes, already allocated resources are not optimized. Either they are working on low priority projects, rather than the type of projects which meet organizational goals - or else they are doing 'busy work'; in other words, they are not assigned challenging tasks according to their skill sets and responsibilities. A project manager needs to be aware of the project development and be able to spot this.

DATA ANALYTICS

Data Analytics in the most basic definition is the science of analysing raw data in order to derive a conclusion about that dataset. The majority of techniques and processes of data analytics have been automated into mechanical processes and algorithms. These techniques can help identify and reveal trends and metrics that might get lost in the mass of information. This new information can then be used to increase the overall efficiency of a business or system.

Data Analytics has 4 main types:

1. **Descriptive analytics:** Tells us what has happened over a particular period of time.
2. **Diagnostic analytics:** This focuses more on varied data inputs and a bit of hypothesizing.
3. **Predictive analytics:** Its objective is to figure out what is possibly going to happen in the near term.
4. **Prescriptive analytics:** This suggests a course of action.

There are a few technologies that have made data analytics all the more powerful. A few of them are:

1. **Machine Learning:** ML is a subset of Artificial Intelligence that is important for Data Analytics and involves algorithms that can learn on their own. ML has the ability to enable applications to take in data and analyse it to predict outcomes without someone explicitly programming the system to reach that conclusion.
2. **Data Management:** It is important to have a procedure in place for easy flow of data in and out of systems and to keep it organised prior to analysing it. Establishing a data management program can help ensure that the organisation is on the same page regarding how to organise and handle data.
3. **Data Mining:** The term data mining refers to the process of sorting through a large amount of data to identify patterns and discover relationships between data points.
4. **Predictive Analytics:** Predictive Analytics technology helps analyse past data to predict future outcomes. These technologies use statistical algorithms and machine learning.

Tools used in Data Analytics

- R Language/R Studio
- Python
- Scala
- Julia
- SAS
- SQL consoles
- Power BI
- Tableau
- MATLAB

PREDICTIVE ANALYTICS

This Analytics type makes use of statistics and modelling techniques to make predictions about future values and outcomes. Predictive analytics looks at current and historical data to determine future performance. Predictive models help make weather forecasts, sales forecasts, customer service decisions, develop video games. Predictive analytics and machine learning are often confused but are two different disciplines. This is used as a decision-making tool in a variety of industries.

Uses of Predictive Analytics

1. **Forecasting:** is an essential part of the manufacturing process as it helps to make sure that resources are used to their full potential in a supply chain. Predictive modelling is often used to clean the data used for such forecasts.
2. **Credit:** Credit scoring makes extensive use of predictive analytics. When a consumer or business applies for credit, data on the applicant's credit history and the credit record of borrowers with similar characteristics are used to predict the risk that the applicant might fail to perform on any credit extended.
3. **Underwriting:** Insurance companies examine policy applicants to determine the likelihood of having to pay out for a future claim based on the current risk pool of similar policyholders, as well as past events that have resulted in payouts.
4. **Marketing:** Individuals who work in this field can analyse and look at how many consumers have reacted to the when planning on a new campaign.

FORECASTING

Forecasting is one of the predictive analytics techniques used to predict future results based on previous data. It uses statistical tools and techniques and is therefore also called statistical analysis. To attain the most advantage from forecasts, the organisations must know the various models and methods that exist.

Prediction and Forecasting are often confused with each other but are different. Prediction is the process of estimating the outcomes of unseen data whereas forecasting is a sub-discipline of prediction in which time-series data is used to make forecasts about the future.

Importance of forecasting:

1. Informs about the challenges of future events.
2. Helps in the management of uncertainty in a better manner.
3. Helps in optimising the use of resources and capital.

Types of forecasting methods:

1. **Qualitative Methods**- In the absence of historical data, qualitative forecasting techniques are sufficient. They are subjective based on the opinion and judgement of the consumers and subject matter experts.
2. **Quantitative Methods**-Forecasting future data as a result of historical data is usually done using quantitative forecasting.
3. **Average Method**-All future values are forecasted to be equal to the mean of the previous data
4. **Naïve Method**- The previous actuals are used as a projection for this period without any adjustments to identify causal factors.
5. **Drift Method**-Allowing predictions to rise or decrease over time is a variant on the naïve process with the amount of change over time fixed to the average change observed in the historical records.

TIME SERIES FORECASTING

Time Series forecasting is a technique for predicting future values by analysing past trends, by assuming that the future trends will consist of similar trends as the historical data. Forecasting involves using models fit on historical data to predict future values. Prediction problems that involve a time component require time series forecasting, which provides a data-driven approach to effective and efficient planning. Historical data is analysed to check for patterns like trends, seasonality etc. These patterns help the data analysts to decide on the model that should be used for predictive modelling.

There are different time series forecasting models such as:

1. Naive model
2. Seasonal Decomposition
3. Exponential Smoothing
4. ARIMA, SARIMA
5. GARCH
6. TBATS
7. Prophet
8. NNETAR

In this Project we have used Exponential Smoothing model.

EXPONENTIAL SMOOTHING

Exponential smoothing is one of the more powerful and common forecasting methods. It produces a forecast based on the weighted average of past observations. In simple terms, these models produce forecasts where the values bear a resemblance to recent observations. Exponential Smoothing allows for weighted averages where greater weight can be placed on recent observations and lesser weight on older observations. Exponential Smoothing is a widely used technique as they are very effective predictors and can be applied to a wide variety of data and use cases.

Exponential smoothing is a preferred method for forecasting in agricultural areas. It is simple to understand, easy to implement with a numerical program and reliable forecast in a wide variety of applications. Some common types of ES consist of:

- Single exponential smoothing (SES)
- Double exponential smoothing (DES)
- Triple exponential smoothing (TES)

Some models are:

1. Simple Exponential Smoothing
2. Holt's Method
3. Holt-Winters Seasonal Method
4. Damped Trend Method

We have specifically used the Holt's Method to predict the crop production.

HOLT'S METHOD

This is a method that works with data having a trend but no seasonality. To make predictions on the data, Holt's Method uses two smoothing parameters which are alpha and beta. Alpha and beta correspond to the level component and trend components. In R, to apply holt's method, the `holt()` function is used. If the alpha and beta values are not set manually, then the `holt()` function will identify the optimum value automatically. The model is trained on actual/observed values collected by the user and then the forecasted values are given as the output.

Holt's model deals with three separate equations that work together to generate a final forecast. The first equation is a basic smoothing equation that adjusts the last smoothed value for the last period's trend. The trend itself is updated through the second equation, where the trend is expressed as the difference between the last two smoothed values. Finally, the third equation is used to generate the final forecast.

LANGUAGES AND TOOLS

R Programming

R is a programming language and software environment used for statistical analysis, graphics representation and reporting. R was created by Ross Ihaka and Robert Gentleman at the University of Auckland, New Zealand. This programming language was named R, based on the first letter of first name of the two R authors (Robert Gentleman and Ross Ihaka), and partly a play on the name of the Bell Labs Language S.

Pros of R Programming:

- Open source
- Platform Independent
- Machine Learning Operation: R allows us to do various machine learning operations such as classification and regression.
- Exemplary support for data wrangling: R provides packages such as dplyr, readr which are capable of transforming messy data into a structured form.
- Quality plotting and graphing: R is a better tool for graphing with packages such as ggplot2 and plotly.
- The array of Packages: R has a rich set of packages. R provides packages for data science and machine learning operations.
- Statistics
- Continuously Growing: R is a constantly evolving programming language.

Power BI

Power BI is a powerful Business Intelligence and Data Visualization tool used to create interactive dashboards using the data present. It has several versions like Desktop, Service-based (SaaS), and mobile Power BI apps. It provides multiple software connectors and services for business intelligence.

Advantages of Power BI:

1. Pre-built dashboards and reports for SaaS Solutions
2. Power BI allows real-time dashboard updates.
3. Offers Secure and reliable connection to your data sources in the cloud or on-premises
4. Power BI offers Quick deployment, hybrid configuration, and secure environment.
5. Allows data exploration using natural language query
6. Offers feature for dashboard visualization regularly updated with the community.

DATABASE OF FARMS

Andhra Pradesh is one of the leading states in India for tobacco production. In this project, we have considered 20 different farms situated in that state with varying sizes and soil types. There are 5 main soil types in Andhra Pradesh for tobacco production:

1. **Northern Black Soil:** The crop is generally planted in October and harvested in Dec-Feb. This soil has a high yield of **1500-1700 Kgs per hectare**. The tobacco produced is good bodied with lengthy leaf, lemon-orange flashy in colour with medium nicotine.
2. **Central Black Soil:** The crop is planted in Oct-Nov and harvested in Dec-Feb. The crop is grown under rainfed conditions and yields are up to **1500 to 1650 Kgs/Ha**. The tobacco produced in this tract is almost similar to that of Northern black soils.
3. **Southern Black Soil:** The crop is planted in Oct-Nov and harvested in Dec-Feb. The average yields are around **1500 Kgs/ha**. The tobacco produced is lemon to lemon-orange in colour and medium-bodied. This tobacco is rated better in quality than the NBS/CBS tobaccos.
4. **Southern Light Soil:** The crop is planted in September-October and harvested in Dec-Jan. The yields are around **1000 kgs/ha**. Erratic rainfall and drought situations generally affect yields in this zone. The leaf is lemon to lemon-orange, thin to medium-bodied, ripe and open grain with good aroma and low in nicotine.
5. **Northern Light Soil:** The crop is planted in Sept-Oct and harvested in Dec-Feb. The yields are ranging between **1800-2000 kgs/ha** depending on the stage of topping. The leaf is medium to heavy-bodied, ripe open-grained, orange to deep orange in colour and fluffy.

The farms based on their size are divided into 4 categories:

1. **Small:** 0.5-2 hectare
2. **Semi-Medium:** 2-4 hectare
3. **Medium:** 4-10 hectare
4. **Large:** 10- above hectare

5 farms belonging to each size are taken where each farm belongs to a different soil.

Therefore, by taking every single case possible here, a database of 20 farms has been created in MS-Excel.

Note: All the data taken in this project is purely for the case study and bears no resemblance/ does not have anything to do with GPI's d

The first sheet of the database workbook consists of a consolidated data of the 20 farms. The variables in this sheet are:

1. Name of the Farm
2. Size (in hectare)
3. Soil (NBS/CBS/SBS/SLS/NLS)
4. Size (small/semi-med/med/large)
5. Production in 2013 (in kgs)
6. Production in 2014 (in kgs)
7. Production in 2015(in kgs)
8. Production in 2016(in kgs)
9. Production in 2017(in kgs)
10. Production in 2018(in kgs)
11. Production in 2019(in kgs)
12. Average Production (in kgs)
13. Average cost of cultivation (in Rs)

Farm Name	Size(hectares)	Soil	Size	2013(in kgs)	2014(in kgs)	2015(in kgs)	2016(in kgs)	2017(in kgs)	2018(in kgs)	2019(in kgs)	Avg production(in kgs)	Avg CostofCultivation(in Rs)
Siri farms	0.94	NBS	small	1,410	1,457	1,504	1316	1541.6	1269	1457	1,422	44,457
Vasant farms	1.2	CBS	small	1,800	1,890	1,922.40	1830	1860	1740	1950	1,856	54,461
Uma farms	1.6	SBS	small	2,360	2,240	2320	2384	2360	2080	2320	2,295	59,747
VSE farms	1.1	NLS	small	1,941.50	2,002	2057	2117.5	2079	1870	2090	2,022.43	59,905
Surva farms	0.8	SLS	small	784	680	760	780	792	640	704	734	34,984
Rao farms	2.5	NBS	semi-med	3,750	3,857	4,000	3500	4125	3375	3875	3,783	70,661
Krishna farms	2.8	CBS	semi-med	4,340	4,410	4488.4	4270	3875	4060	4550	4,285	77,429
DhanaRai farms	3.5	SBS	semi-med	5,162.60	4,900	5075	5215	5162.5	4550	5075	5,020.01	81,341
Raju farms	3.2	NLS	semi-med	5,648	5,842	5984	6160	6048	5440	6080	5,886	84,345
Devaki farms	3.7	SLS	semi-med	3,626	3,145	3515	3607.5	3663	2960	3256	3,396	72,837
Lakshmi farms	4.2	NBS	med	6,300	6,510	6,720	5880	6930	5670	6510	6,360	85,502
Hemalatha farms	4.7	CBS	med	7285	7,402.50	7529.4	7167.5	7285	6815	7637.5	7,303	96,860
Bindu farms	5.6	SBS	med	8260	7,840	8120	8344	8260	7280	8120	8,032	1,00,550
Vasundra farms	5.9	NLS	med	10,413.50	10,738	11033	11357.5	11151	10030	11970	10,956.14	1,14,215
Rama farms	6.3	SLS	med	6,174	5,355	5985	6142.5	6237	5040	5544	5,783	79,722
Rani Farms	10.1	NBS	large	15,150	15,655	16,160	14140	16665	13635	15810	15,316	1,17,123
Prt Ltd farms	11	CBS	large	17,050	17,325	17622	16775	17050	15950	17875	17,092	1,21,245
RR farms	10.3	SBS	large	15,192	14,420	14935	15347	15192.5	13390	14935	14,773	1,10,120
Sanch farms	10	NLS	large	17,650	18,200	18700	19250	18900	17000	19000	18,386	1,22,100

Figure 1: The main consolidated sheet

Further, a separate sheet in excel is created for each farm. The variables taken into account for each farm were:

1. Year
2. Tobacco Crop Production for years 2013-2019
3. Labour Cost (in Rs)
4. Seed Cost (in Rs)
5. Fertilizer Cost (in Rs)
6. Fixed Cost (in Rs)
7. Total Cost (Sum of Labour, Seed, Fertilizer and Fixed cost) (in Rs)
8. fprod (Forecasted production)
9. flab (Forecasted labour cost)
10. fseed (Forecasted Seed cost)
11. ffert (Forecasted Fertiliser cost)
12. ffixed (Forecasted Fixed cost)
13. ftotal (Forecasted Total cost)

Note: The unit in which the production is measured is kilograms(kgs) and the unit in which the cost if being taken is Rupees (Rs) throughout this project.

Year	production	fprod	Labour cost	flab	Seed Cost	fseed	Fertiliser Cost	ffert	Fixed Cost	ffixed	Total cost	ftotal
01-08-2013	2360		14,500		35,000		4000		5,000		58,500	
01-08-2014	2240		14,000		35,200		4500		5,500		59,200	
01-08-2015	2320		14,000		36,300		5000		5,000		60,300	
01-08-2016	2384		14,500		36,000		6000		5,000		61,500	
01-08-2017	2360		14,500		35,550		5100		5,000		60,150	
01-08-2018	2080		14,600		34,400		4000		5,500		58,500	
01-08-2019	2320	2320	15,000	15,000	35,400	35,400	4000	4000	5,000	5,000	59,400	59,400
01-08-2020		2245.046		15,207		35,500		3843.117		5,129		59,680
01-08-2021		2439.22		15,470		35,705		3686.233		5,106		59,968
01-08-2022		2333.413		15,733		35,905		3529.35		5,100		60,268

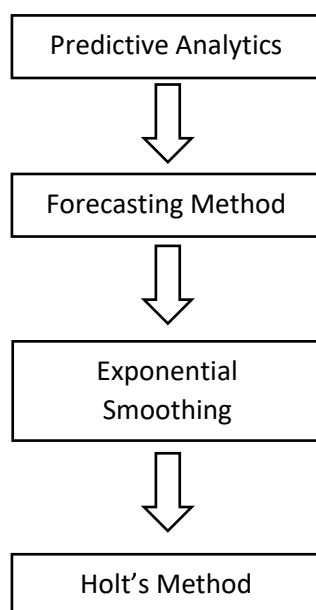
Figure 2: The excel sheet for a single farm

The costs that are taken into account are:

1. **Seed Cost:** The cost of Seeds and Seedlings
2. **Labour Cost:** Labour Cost includes the following costs
 - Human labour
 - a. Family Labour
 - b. Hired Labour
 - Bullock Labour
 - a. Owned
 - b. Hired
 - Tractor
 - a. Owned
 - b. Hired
 - Transportation
3. **Fertilizer Cost:** Fertilizer cost includes the cost of manures and fertilisers
 - Manures and Fertilizer
 - Plant Protection Chemicals
 - Fuel Wood
4. **Fixed Cost:** Fixed cost consists of a total of costs incurred from
 - Land Revenue
 - Tobacco Board license fee
 - Depreciation
 - Rental value of owned land
 - Interest on fixed capital

FORECASTING DATA

As mentioned earlier, Predictive Analysis is used to determine crop production for the next 3 years i.e. 2020,2021 and 2022.



The value observed value from 2013-2019 have been used to create the database. The following variables are being predicted for each farm for the years 2020,2021 and 2022:

1. Production
2. Labour Cost
3. Seed Cost
4. Fertilizer Cost
5. Fixed Cost

Exponential Smoothing is the method being used, as in this method, the more weights are given to the recent observations and lesser weight to past observations. This is a suitable method to predict these values as over a period of time as the cost keeps on changing for seeds, fertilisers, labour and land. Along with this, the changes in the soil also affects the production over a time frame. Under exponential smoothing, the Holt's Method is used for forecasting as there is a trend in production and the various costs taken into account for the farms.

The model is being implemented using R programming language using R studio.

The R-packages used:

1. **Tidyverse:** The packages under the tidyverse umbrella help us in performing and interacting with the data. There are a whole host of things you can do with your data, such as subsetting, transforming, visualizing, etc.
2. **Dplyr:** dplyr is a package for making data manipulation easier.
3. **Forecast:** Provides methods and tools for displaying and analysing univariate time series forecasts including exponential smoothing via state space models and automatic ARIMA modelling.
4. **fpp2:** It is also a package that contains forecasting methods.
5. **TTR:** TTR is an R package that provides the most popular technical analysis functions
6. **fma:** Forecasting: methods and applications
7. **ggplot2:** ggplot2 is a powerful and a flexible R package used for producing elegant graphics.
8. **Readxl:** The readxl package makes it easy to get data out of Excel and into R.

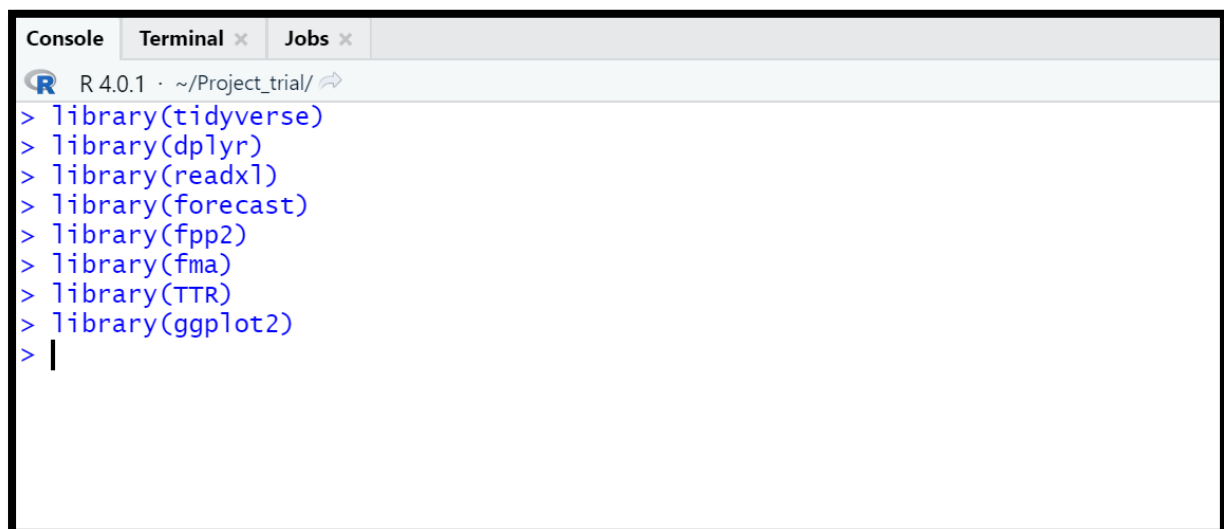
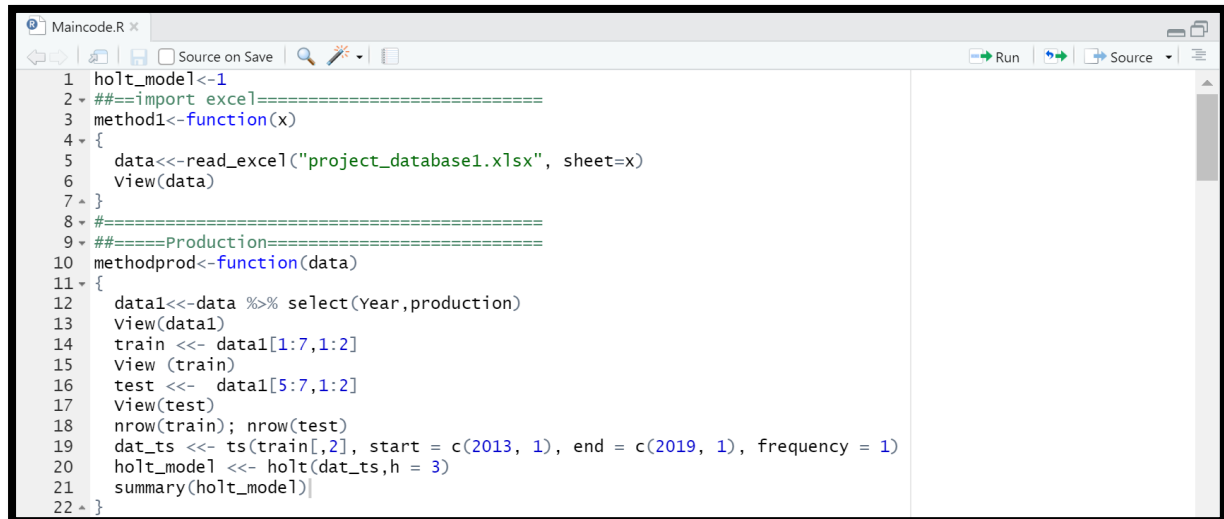
A screenshot of the R Studio interface, specifically the Console pane. The title bar shows 'Console', 'Terminal', and 'Jobs'. The console output shows the R prompt followed by the loading of several packages: library(tidyverse), library(dplyr), library(readxl), library(forecast), library(fpp2), library(fma), library(TTR), and library(ggplot2). The prompt is currently at the line '> |'.

Figure 3: Loading required Packages

The holt() function is used by holt method which is a part of forecast and fpp2 package to forecast values. Here we have not taken alpha and beta values in holt() function manually, therefore the holt() function will identify the optimum value automatically.

The code used to predict values:



```

1 holt_model<-1
2 ##==import excel=====
3 method1<-function(x)
4 {
5   data<-read_excel("project_database1.xlsx", sheet=x)
6   view(data)
7 }
8 ##=====Production=====
9
10 methodprod<-function(data)
11 {
12   data1<-data %>% select(Year,production)
13   view(data1)
14   train <- data1[1:7,1:2]
15   view (train)
16   test <- data1[5:7,1:2]
17   view(test)
18   nrow(train); nrow(test)
19   dat_ts <- ts(train[,2], start = c(2013, 1), end = c(2019, 1), frequency = 1)
20   holt_model <- holt(dat_ts,h = 3)
21   summary(holt_model)
22 }

```

Figure 4: Code to import Excel sheet and to predict production

1. Method 1 contains read_excel () function which is used to import the sheet of the particular farm from Excel into R studio. A sheet can be loaded into R studio by writing the sheet name inside method1("Name") function while calling it.
2. Methodprod () is a function created. The code within this function allows the user to forecast the production for the years 2020,2021 and 2022.
The Year and the Production column are selected using select () function and saved in data1.
Using the values in data1, the dataset is divided into train and test set.
3. The train table is converted into time series data and saved in dat_ts.
4. Holt method is applied to dat_ts. Here h=3 refers to the frequency.
h=3 gives the forecasted values for next 3 years.
5. Summary () function gives the forecasted values.

```

inside1<-function()
{
  mape <-> function(actual,pred)
  {
    mape <-> mean(abs((actual - pred)/actual))*100
    return (mape)
  }

  df_holt <-> as.data.frame(holt_model)
  test$holt <-> df_holt$`Point Forecast`
  mape(test$production,test$holt)
}

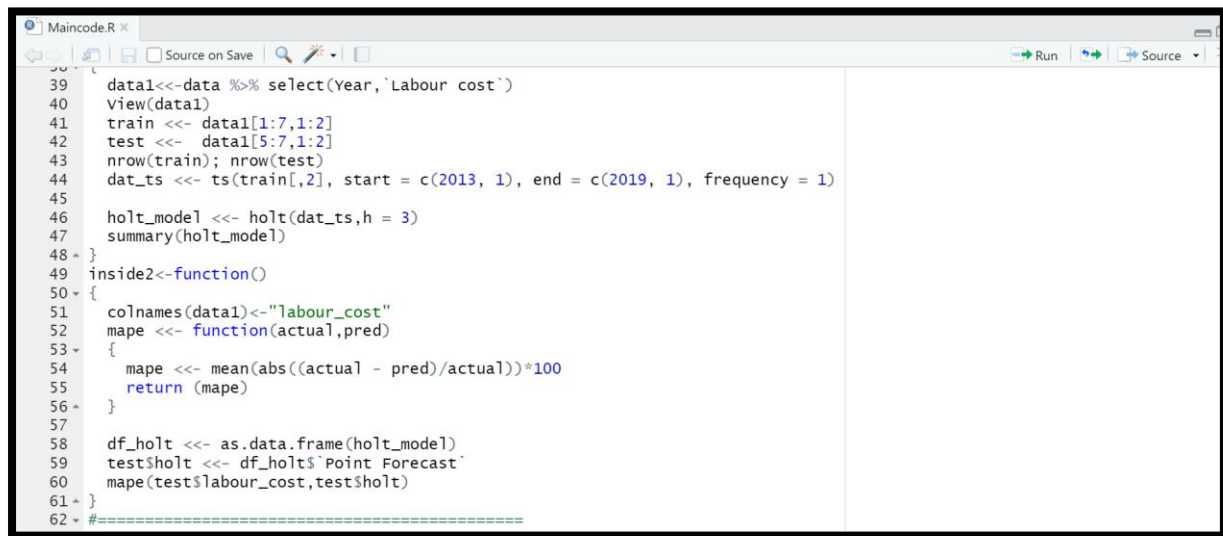
```

Figure 5: The inside() function to calculate error

6. The inside () function has the code which allows the user to calculate the MAPE(Mean Absolute Percentage Error).
 MAPE: is a statistical error metric that estimates the accuracy of a model for the input dataset. Using MAPE, one can understand the difference between the actual and the predicted values in terms of misclassification as well as accuracy.
 It is usually considered that lower the MAPE, better fit is the applied model.

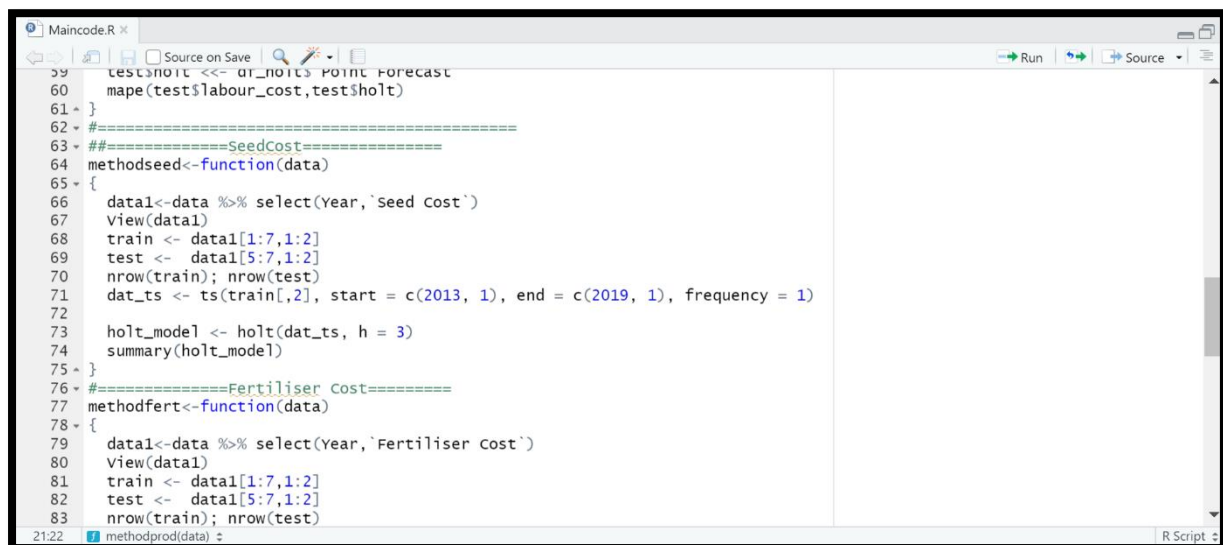
Similarly, the variable production is replaced in the code in Figure 4 and Figure 5 with required variables such as labour cost, seed cost, fertiliser cost and fixed cost to forecast these values.

For instance,



```
39 data1<-data %>% select(Year,'Labour cost')
40 view(data1)
41 train <- data1[1:7,1:2]
42 test <- data1[5:7,1:2]
43 nrow(train); nrow(test)
44 dat_ts <- ts(train[,2], start = c(2013, 1), end = c(2019, 1), frequency = 1)
45
46 holt_model <- holt(dat_ts,h = 3)
47 summary(holt_model)
48 }
49 inside2<-function()
50 {
51   colnames(data1)<-"labour_cost"
52   mape <- function(actual,pred)
53   {
54     mape <- mean(abs((actual - pred)/actual))*100
55     return (mape)
56   }
57
58   df_holt <- as.data.frame(holt_model)
59   test$holt <- df_holt$`Point Forecast`
60   mape(test$labour_cost,test$holt)
61 }
62 #=====
```

Figure 6: Code to forecast Labour cost and calculate its MAPE



```
59 test$holt <- df_holt$`Point Forecast`
60 mape(test$labour_cost,test$holt)
61 }
62 #=====
63 #=====SeedCost=====
64 methodseed<-function(data)
65 {
66   data1<-data %>% select(Year,'Seed Cost')
67   view(data1)
68   train <- data1[1:7,1:2]
69   test <- data1[5:7,1:2]
70   nrow(train); nrow(test)
71   dat_ts <- ts(train[,2], start = c(2013, 1), end = c(2019, 1), frequency = 1)
72
73   holt_model <- holt(dat_ts, h = 3)
74   summary(holt_model)
75 }
76 #=====Fertiliser Cost=====
77 methodfert<-function(data)
78 {
79   data1<-data %>% select(Year,'Fertiliser cost')
80   view(data1)
81   train <- data1[1:7,1:2]
82   test <- data1[5:7,1:2]
83   nrow(train); nrow(test)
84 }
21:22 methodprod(data)
R Script
```

Figure 7: Code to forecast seed cost and fertiliser cost

The output shown on executing these functions:

```

105 mape <- function(actual, pred)
106 {
107     mape <- mean(abs((actual - pred)/actual))*100
108     return (mape)
109 }
110
111 df_holt <- as.data.frame(holt_model)
112 test$holt <- df_holt$`Point Forecast`
113 mape(test$production, test$holt)
114 }
115
116 method1("KKM Farm")
117 methodprod(data)
118 inside1()
119 methodlab(data)
120 inside2()
121 methodseed(data)
122 inside3()
123 methodfert(data)
124 inside4()
125 methodfixed(data)
126 inside5()
127
128
129

```

Figure 8: Calling the functions

1. By replacing the name of the required sheet in the double quotes inside the method1() function, that particular sheet containing the data about the particular farm will be loaded. This is called calling the function method1().
2. Once the sheet is loaded it is saved into the variable called data, which is a global variable. The data is now being given as an argument in each function select the required columns and perform predictive analysis.
3. User only need to change the name and run each of these lines of code and they will have the desired output.

```

Forecasts:
  Point Forecast   Lo 80   Hi 80   Lo 95   Hi 95
2020  8534.584 6903.899 10165.27 6040.666 11028.50
2021  8297.796 5744.941 10850.65 4393.541 12202.05
2022  8061.008 4320.288 11801.73 2340.071 13781.94
> inside1()
[1] 9.121623
>

```

Figure 9: The forecasted values for production

Here there is an approximately 9% error between the predicted and the test values.

Similarly, each of the functions for seed cost, labour cost, fertiliser cost and fixed cost has to be run in order to get their forecasted values.

CREATING VISUALISATIONS

Power BI is a strong data visualisation tool that is used by several major organisations to help with data analytics. In this project, several graphs have been created to draw relationships and observe trend to draw conclusion.

These visualisations have been created using the actual data from 2013-2019.

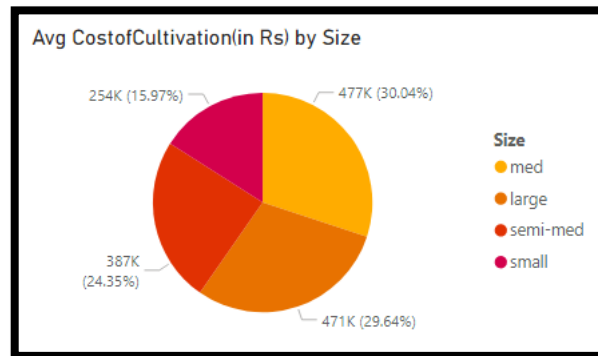


Figure 10: Pie chart showing Average Cost of cultivation by Size

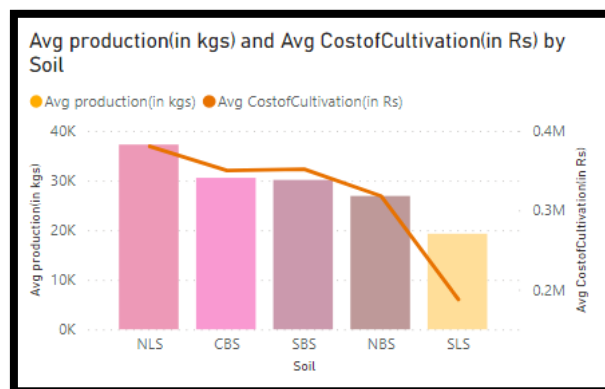


Figure 11: Line and column chart showing Average production and Average cost of cultivation by Soil

From these visualisations, one can see that Northern Light soil (NLS) has the highest average production, indicating that it is the most fertile and optimum soil for tobacco production.

Similarly, separate dashboards have been created for each and every farm. Each dashboard consists of the following graphs: -

1. A line and stacked column chart presenting **Production and Total Cost by Year**.
2. A line and clustered column chart presenting **Seed cost, Labour cost, Fertiliser Cost, Fixed Cost and Total Cost by Year**.

The above charts are created using the actual data from 2013-2019.

3. A line chart showing the **actual production and the projected Projection(fprod) by Year** which is represented by a solid line and a dotted line respectively.
4. A line chart showing the **actual Total Cost and the projected Total Cost(ftotal) by Year** which is represented by a solid line and a dotted line respectively.
5. A multiple line chart showing the **actual Seed cost, Labour cost, Fertiliser Cost, Fixed Cost and the projected Seed cost (fseed), Labour cost(flabor), Fertiliser Cost(ffert), Fixed Cost (ffixed) by Year** which is represented by a solid line and a dotted line respectively.

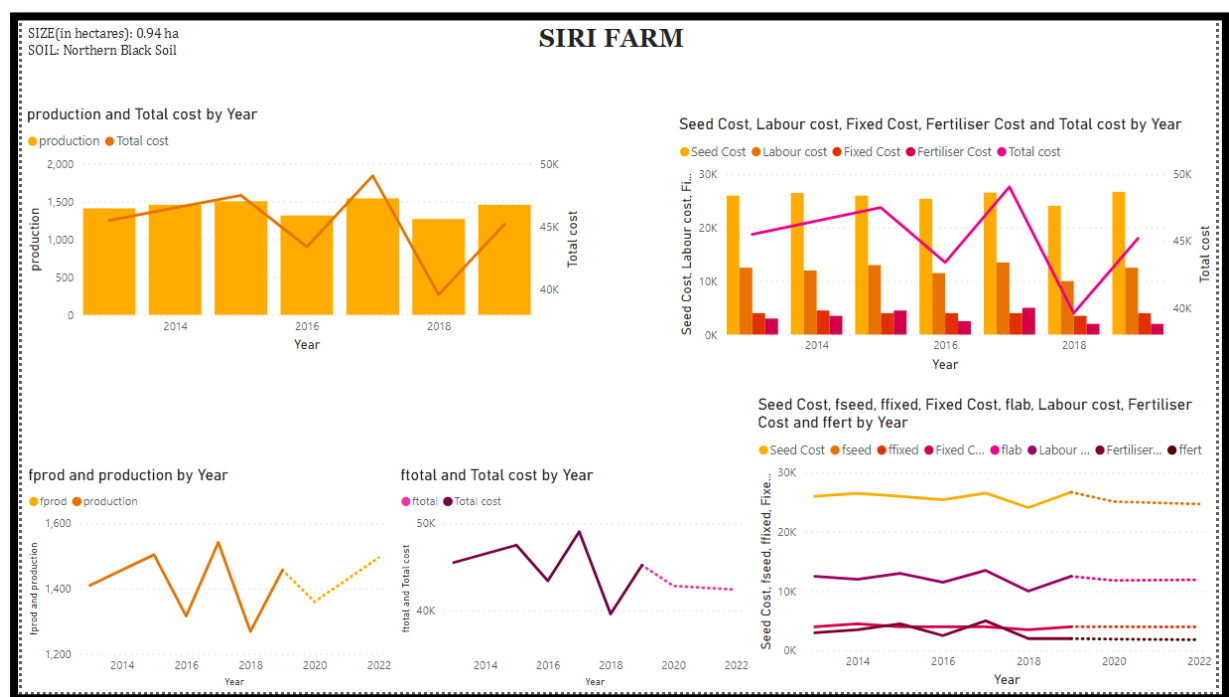


Figure 12: Dashboard of a farm

CONCLUSION

In this project, after forecasting the required values for each farm, their dashboards were created to analyse the trend. It was noticed that the projected production went up while the projected total cost went down for 50% of the farms. It was seen that the production went up while the total cost went down which shows a positive conclusion where production increases for less cost.

For about 20% of the farms, it was observed that as the forecasted production went down, the forecasted total cost went down as well. There is a negative slope in production in this case.

For 70% of the farms, the Holts model helped us forecast values which resulted in a more desirable outcome. The main aim is to reduce total costs and make sure these resources are used to their maximum potential. There is an efficient use of the resources provided as the predicted production is either increasing or decreasing with decreasing or lesser resources.

The rest of the 30% of farms, did not show a positive outcome as in these cases the production decreased or increased with an increase in total cost. Through better resource allocation in the future, the total cost and other resources can be optimised in these farms.

Using this data, the organisation can make decisions regarding investing in these farms and paying for the cultivation of tobacco. The organisation will now have an approximate idea of how much tobacco will be produced and its cultivation cost.

This table is a consolidated version of all the farms:

Farm Name	Production (in kgs)				Cost of Cultivation (in Rs)			
	Avg. Production from 2013-2019	Prod. in 2020	Prod. in 2021	Prod. in 2022	Avg. cost of cultivation from 2013-2019	Cost of cult. in 2020	Cost of cult. in 2021	Cost of cult. in 2022
Siri Farm	1,422	1,360	1,427	1,495	44,457	42,823	42,594	42,406
Vasant Farm	1,856	1,851	1,959	1,967	54,461	55,603	56,505	56,449
Uma Farm	2,295	2,245	2,439	2,333	59,747	59,680	59,968	60,268
VSF Farm	2,022.43	2,014	1,982	1,949	59,905	60,665	60,251	59,837
Surya Farm	734	669.37	650.8	632.235	34,984	33,415	32,479	31,542
Rao Farm	3,783	3,657	3,776	4,495	70,661	70,081	70,499	70,933
Krishna Farm	4,285	5,040	5,530	6,020	77,429	78,872	78,813	78,698
Dhanraj Farm	5,020	4,911	4,898	5,385	81,341	83,013	81,366	81,727
Raju Farm	5,886	5,825	5,721	5,618	84,345	86,404	84,646	84,102
Devaki Farm	3,396	3,095	3,009	3,224	72,837	72,782	70,894	69,396
Lakshmi Farm	6,360	6,096	5,950	6,003	85,502	84,743	85,751	85,029
Hemalatha Farm	7,303	7,182	7,129	7,576	96,860	97,174	96,563	97,058
Bindu Farm	8,032	7,857	7,837	7,816	1,00,550	1,00,931	97,366	98,605
Vasundra Farm	10,956	11,348	11,346	11,353	1,14,215	1,16,860	1,16,730	1,18,378
Rama Farm	5,783	5,271	5,125	5,567	79,722	79,137	78,798	79,386
Rani Farm	15,316	14,753	14,431	14,510	1,17,123	1,16,359	1,15,511	1,14,663
Prt Ltd Farm	17,092	17,809	16,985	17,900	1,21,245	1,20,965	1,18,714	1,20,170
RR Farm	14,773	14,452	14,615	14,756	1,10,120	1,10,637	1,10,488	1,09,234
Sanch Farm	18,386	18,317	18,022	18,500	1,22,100	1,21,912	1,21,668	1,21,424
KKM Farm	9,362	8,534	8,297	8,061	1,04,875	1,03,830	1,03,385	1,02,939

Figure 13: Conclusion Table

	Projected production increased while the projected total cost decreased
	Projected production decreased while the projected total cost decreased
	Projected production increased or decreased while the projected total cost increased

SUMMARY

The main aim of this project was to optimise the resources put into farms by predicting the production of tobacco for the next 3 years using data analytics. Production and cost of cultivation from 2013-2019 was taken and the same variables were predicted for the next three years i.e., 2020 to 2022. Holt's method which is a model under exponential smoothing was used. Exponential smoothing is a forecasting method under predictive analytics. The database was created in excel and the forecasting model was implemented using R programming in R studio. After forecasting, this data was represented using line and bar graphs in Power BI for each farm to analyse the trend.

Resource optimisation happens when the maximum potential is observed when the cost incurred is less. It was observed that 70% of the farms had an increasing or decreasing projected production with a **decreasing** cost of cultivation which is a desirable outcome. While 30% of the farms had an increasing or decreasing projected production with an **increasing** cost of cultivation.

RECOMMENDATION

I believe that resource optimisation and cutting costs are some of the most important parts to keep the company in profit. Data Analytics and Machine Learning have given companies a variety of tools to analyse their data and make much more informed business decisions than before. Forecasting techniques in Data Analytics is not only one of the most common but also a very reliable method to predict values. The cost of cultivation is usually provided by the company to farmers who produce tobacco for their company. By predicting the crop production and its cost, it would allow them to invest as many resources as needed.

Therefore, I would recommend this method or a similar one to predict crop production. The outcome that matters is that the farms' maximum production capacity is reached with a smaller number of resources. By using larger datasets and better models, more accurate values can be obtained and more informed decisions can be made.

Data Analytics is a very broad topic that contains tools to help people make better business decisions. Other than this, technologies like Artificial Intelligence(AI), Internet of Things (IoT), Machine Learning (ML) have a wide scope of improving and automating processes and calculations that usually take manual power or more time.

REFERENCES

1. [https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp#:~:text=Fast-moving consumer goods are, products%2C and baked goods\).](https://www.investopedia.com/terms/f/fastmoving-consumer-goods-fmcg.asp#:~:text=Fast-moving consumer goods are, products%2C and baked goods).)
2. <https://tobaccoboard.com/soiltypes.php>
3. <http://oaji.net/articles/2016/491-1478776171.pdf>
4. http://mospi.nic.in/sites/default/files/publication_reports/manual_cost_cultivation_surveys_23july08_0.pdf
http://mospi.nic.in/sites/default/files/publication_reports/manual_cost_cultivation_surveys_23july08_0.pdf
5. http://mospi.nic.in/sites/default/files/publication_reports/manual_cost_cultivation_surveys_23july08_0.pdf
http://mospi.nic.in/sites/default/files/publication_reports/manual_cost_cultivation_surveys_23july08_0.pdf
6. <https://www.lotame.com/what-is-data-analytics/>
7. <https://www.lotame.com/what-is-data-analytics/>
8. <https://www.onemodel.co/blog/ai-academy-forecasting-vs-predictive-modeling>
9. <https://www.xenonstack.com/insights/what-is-forecasting>
10. <https://www.investopedia.com/terms/p/predictive-analytics.asp>
11. <https://www.infoworld.com/article/3622246/an-introduction-to-time-series-forecasting.html>
12. <https://www.tutorialspoint.com/r/index.htm>
13. <https://www.javatpoint.com/r-advantages-and-disadvantages>
14. <https://www.guru99.com/power-bi-tutorial.html>
15. <https://www.researchandmarkets.com/reports/4757741/tobacco-market-in-india-2018-2023>
16. (2000) HOLT'S FORECASTING MODEL. In: Swamidass P.M. (eds) Encyclopedia of Production and Manufacturing Management. Springer, Boston, MA . https://doi.org/10.1007/1-4020-0612-8_409
17. **Great Learning Team** (2020). *How Machine Learning is Simplifying Sales Forecasting & Increasing Accuracy*, Feb 14, 2020