**Power BI enabled crop production analysis**

**College Name**

**Government arts and science college**

**vedaranyam**

|  |  |
| --- | --- |
| **NM ID** | **NAME** |
| asbdu369cb21s206504 | E.Mohana |

Trainer name

Uma maheswari

**ABSTRACT**

Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research.In this study, we performed a Systematic Literature Review (SLR) to extract and synthesize the algorithms and features that have been used in crop yield prediction studies. Based on our search criteria, we retrieved 567 relevant studies from six electronic databases, of which we have selected 50 studies for further analysis using inclusion and exclusion criteria. We investigated these selected studies carefully, analyzed the methods and features used, and provided suggestions for further research.According to our analysis, the most used features are temperature, rainfall, and soil type, and the most applied algorithm is Artificial Neural Networks in these models. After this observation based on the analysis of machine learning-based 50 papers, we performed an additional search in electronic databases to identify deep learning -based studies, reached 30 deep learning-based papers, and extracted the applied deep learning algorithms. According to this additional analysis,Convolutional Neural Networks (CNN) is the most widely used deep learning algorithm in these studies, and the other widely used deep learning algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks (DNN).

**INDEX**

|  |  |  |
| --- | --- | --- |
| **Sr. No.** | **Table of Contents** | **Page No.** |
| 1 | Chapter 1: Introduction | 4 |
| 2 | Chapter 2: Services and Tools Required | 6 |
| 3 | Chapter 3: Project Architecture | 7 |
| 4 | Chapter 4: Modeling and Result | 9 |
| 5 | Conclusion | 18 |
| 6 | Future Scope | 19 |
| 7 | References | 20 |
| 8 | Links | 21 |

**CHAPTER 1**

**INTRODUCTION**

* **Problem Statement**

This study showed that the selected publications use a variety of features, depending on the scope of the research and the availability of data. Every paper investigates yield prediction with machine learning but differs from the features. The studies also differ in scale, geological position, and crop. The choice of features is dependent on the availability of the dataset and the aim of the research. Studies also stated that models with more features did not always provide the best performance for the yield prediction. To find the best performing model, models with more and fewer features should be tested. Many algorithms have been used in different studies. The results show that no specific conclusion can be drawn as to what the best model is, but they clearly show that some machine learning models are used more than the others. The most used models are the random forest, neural networks, linear regression, and gradient boosting tree. Most of the studies used a variety of machine learning models to test which model had the best prediction.

* **Proposed Solution**

The proposed solution is to develop a PowerBI dashboard that can analyze and visualize real-time customer data. The dashboard will integrate data from various sources such as transaction history, customer feedback, and demographic data. It will provide a comprehensive view of customer behavior, preferences, and trends, enabling banks to make informed decisions. The dashboard will be interactive, user-friendly, and customizable, allowing banks to tailor it to their specific needs. The real-time analysis capability of the dashboard will enable banks to respond promptly to changes in customer behavior or preferences, identify opportunities for cross-selling and up-selling, and tailor their products and services to meet customer needs.

* **Feature**
* **Real-Time Analysis**: The dashboard will provide real-time analysis of customer data.
* **Customer Segmentation**: It will segment customers based on various parameters like age, income, transaction behavior, etc.
* **Trend Analysis**: The dashboard will identify and display trends in customer behavior.
* **Predictive Analysis**: It will use historical data to predict future customer behavior.
* **Advantages**
* **Data-Driven Decisions**: Banks can make informed decisions based on real-time data analysis.
* **Improved Customer Engagement**: Understanding customer behavior and trends can help banks engage with their customers more effectively.
* **Increased Revenue**: By identifying opportunities for cross-selling and up-selling, banks can increase their revenue.
* **Scope**

The scope of this project extends to all banking institutions that aim to leverage data for decision-making and customer engagement. The project can be further extended to incorporate more data sources and advanced analytics techniques, such as machine learning and artificial intelligence, to provide more sophisticated insights into customer behavior. The project also has the potential to be adapted for other sectors, such as retail, healthcare, and telecommunications, where understanding customer behavior is crucial. Furthermore, the project contributes to the broader goal of digital transformation in the banking sector, promoting efficiency, innovation, and customer-centricity.

**CHAPTER 2**

**SERVICES AND TOOLS REQUIRED**

**2.1 Services Used**

* **Data Collection and Storage Services**: Banks need to collect and store customer data in real-time. This could be achieved through services like Azure Data Factory, Azure Event Hubs, or AWS Kinesis for real-time data collection, and Azure SQL Database or AWS RDS for data storage.
* **Data Processing Services**: Services like Azure Stream Analytics or AWS Kinesis Data Analytics can be used to process the real-time data.
* **Machine Learning Services**: Azure Machine Learning or AWS SageMaker can be used to build predictive models based on historical data.

**2.2 Tools and Software used**

**Tools**:

* **PowerBI**: The main tool for this project is PowerBI, which will be used to create interactive dashboards for real-time data visualization.
* **Power Query**: This is a data connection technology that enables you to discover, connect, combine, and refine data across a wide variety of sources.

**Software Requirements**:

* **PowerBI Desktop**: This is a Windows application that you can use to create reports and publish them to PowerBI.
* **PowerBI Service**: This is an online SaaS (Software as a Service) service that you use to publish reports, create new dashboards, and share insights.
* **PowerBI Mobile**: This is a mobile application that you can use to access your reports and dashboards on the go.

**CHAPTER 3**

**PROJECT ARCHITECTURE**

**3.1 Architecture**



Here’s a high-level architecture for the project:

* **Data Collection**: Real-time customer data is collected from various sources like bank transactions, customer interactions, etc. This could be achieved using services like Azure Event Hubs or AWS Kinesis.
* **Data Storage**: The collected data is stored in a database for processing. Azure SQL Database or AWS RDS can be used for this purpose.
* **Data Processing**: The stored data is processed in real-time using services like Azure Stream Analytics or AWS Kinesis Data Analytics.
* **Machine Learning**: Predictive models are built based on processed data using Azure Machine Learning or AWS SageMaker. These models can help in predicting customer behavior, detecting fraud, etc.
* **Data Visualization**: The processed data and the results from the predictive models are visualized in real-time using PowerBI. PowerBI allows you to create interactive dashboards that can provide valuable insights into the data.
* **Data Access**: The dashboards created in PowerBI can be accessed through PowerBI Desktop, PowerBI Service (online), and PowerBI Mobile.

This architecture provides a comprehensive solution for real-time analysis of bank customers. However, it’s important to note that the specific architecture may vary depending on the bank’s existing infrastructure, specific requirements, and budget. It’s also important to ensure that all tools and services comply with relevant data privacy and security regulations.

**CHAPTER 4**

**MODELING AND RESULT**

**Manage relationship**

The “disp” file will be used as the main connector as it contains most key identifier (account id, client id and disp id) which can be use to relates the 8 data files together. The “district” file is use to link the client profile geographically with “district id”







**Modelling for Gender and Age data**

Notice that the Gender and age of the client are missing from the data. These can be formulated from the birth number YYMMDD where at months (the 3rd and 4th digits) greater than 50 means that client is a Female. We can create a column for Gender.

For birthday, we need to reduce the birth month of the female by 50 and then change the date format to DD/MM/YYYY adding 1900 to the year.

For Age, we shall assume it is year 1999 as explain previously and use it to minus from the birth year.

**Replacing values**

Set some fields to English for easy understanding, we replace values to English with the Power Query Editor.





Changing the order of Region name at Power Query

Duplicate the “district /region” then split column using space as delimiter.

Then merge column by Region and direction. Refer to applied steps for details.

**Grouping of age by ranges**

As the customers’ age ranges from 12 to 88, we shall group them into different generation age range for easier profiling, we will group the ages into 5 groups.

The Gen Y are youths,

Gen X are young working adults, some starting their families

Baby Boomer are working adults with families.

The silent Generations some are working and retired, living on pensions.

The greatest Generation, retired elderly living on pensions.



**Credit Rating and Loan Status**

As the Loan status uses A, B, C, D which are not reader friendly. We can add a column to represent what it stands for, we also simplify the classification of those with late or default on payment as bad credit, refer to the table below for details on the new columns added.

Values of such as “account Id” have also been set as Text.

And District name have been categorized as place to be use for the map to show the sum of the inhabitants in each region.

**Dashboard**







**CONCLUSION**

Machine learning is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. Several machine learning algorithms have been applied to support crop yield prediction research.In this study, we performed a Systematic Literature Review (SLR) to extract and synthesize the algorithms and features that have been used in crop yield prediction studies. Based on our search criteria, we retrieved 567 relevant studies from six electronic databases, of which we have selected 50 studies for further analysis using inclusion and exclusion criteria. We investigated these selected studies carefully, analyzed the methods and features used, and provided suggestions for further research.According to our analysis, the most used features are temperature, rainfall, and soil type, and the most applied algorithm is Artificial Neural Networks in these models. After this observation based on the analysis of machine learning-based 50 papers, we performed an additional search in electronic databases to identify deep learning -based studies, reached 30 deep learning-based papers, and extracted the applied deep learning algorithms. According to this additional analysis,Convolutional Neural Networks (CNN) is the most widely used deep learning algorithm in these studies, and the other widely used deep learning algorithms are Long-Short Term Memory (LSTM) and Deep Neural Networks

**FUTURE SCOPE**

This study showed that the selected publications use a variety of features, depending on the scope of the research and the availability of data. Every paper investigates yield prediction with machine learning but differs from the features. The studies also differ in scale, geological position, and crop. The choice of features is dependent on the availability of the dataset and the aim of the research. Studies also stated that models with more features did not always provide the best performance for the yield prediction. To find the best performing model, models with more and fewer features should be tested. Many algorithms have been used in different studies. The results show that no specific conclusion can be drawn as to what the best model is, but they clearly show that some machine learning models are used more than the others. The most used models are the random forest, neural networks, linear regression, and gradient boosting tree. Most of the studies used a variety of machine learning models to test which model had the best prediction.

**REFERENCES**

<https://www.novypro.com/project/crop-production-in-india-%7C-powerbi-dashboard>