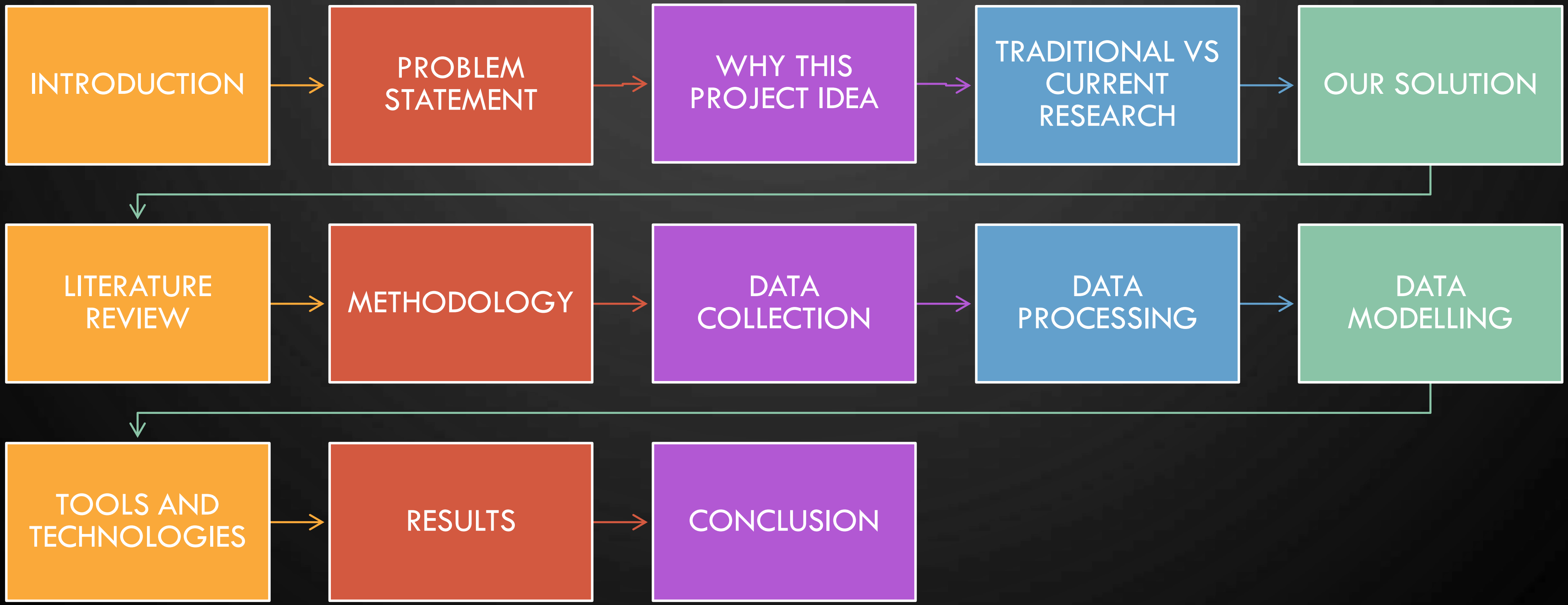
LEVERAGING DATA SCIENCE FOR PREDICTING BREAST CANCER OUTCOMES USING THE WISCONSIN BREAST CANCER DATASET

Presented By: Muhammad Irfan (23173372)
Supervisor: Dr. Abdulrahman Al Sewari

TABLE OF CONTENTS



INTRODUCTION

PROBLEM STATEMENT

WHY THIS PROJECT IDEA

TRADITIONAL VS CURRENT RESEARCH

OUR SOLUTION

LITERATURE REVIEW

METHODOLOGY

DATA COLLECTION

DATA PROCESSING

DATA MODELLING

TOOLS AND TECHNOLOGIES

RESULTS

CONCLUSION

INTRODUCTION

INTRODUCTION

Breast cancer remains one of the most common and deadly forms of cancer among women worldwide. The urgency to improve early detection and accurate prognosis is paramount, as these can significantly influence treatment outcomes and survival rates.

The purpose of this study is to leverage machine learning (ML) models to enhance the prediction accuracy of breast cancer outcomes, specifically using the Wisconsin Breast Cancer Dataset (WBCD). By developing more precise predictive tools, the research aims to contribute to the healthcare field by providing clinicians with better diagnostic support, ultimately leading to improved patient care and reduced healthcare costs.

PROBLEM STATEMENT

PROBLEM STATEMENT

Current methodologies for predicting breast cancer outcomes often suffer from inconsistencies in accuracy and limited generalizability across diverse patient populations.

Traditional statistical models, while useful, have not kept pace with the advancements in data science and machine learning, which have the potential to offer more sophisticated and accurate predictions.

WHY THIS PROJECT IDEA?

WHY THIS PROJECT IDEA?

Breast cancer is a global health concern, and improving the tools available for early detection and accurate prognosis could have a profound impact on patient outcomes.

This project is particularly relevant as it combines the latest advancements in machine learning with a critical healthcare need. By focusing on breast cancer prediction, the research not only contributes to the scientific community but also addresses a public health issue.

The idea to leverage machine learning comes from its proven success in other domains, and its potential to revolutionize cancer diagnosis by providing more accurate, personalized predictions.

TRADITIONAL VS. CURRENT RESEARCH

TRADITIONAL VS. CURRENT RESEARCH

- Traditional approaches to breast cancer prognosis have relied heavily on statistical models, which, while effective in certain contexts, often lack the flexibility and accuracy of modern machine learning techniques. Current research has shifted towards using more advanced algorithms that can handle complex, high-dimensional data.
- This shift has led to improved prediction accuracy and the ability to uncover patterns that were previously undetectable. However, these advancements have also introduced challenges, such as the need for more extensive computational resources and the complexity of model interpretation.
- This study seeks to bridge the gap between traditional and current research by developing models that are both accurate and interpretable.

OUR SOLUTION

OUR SOLUTION

01

To tackle the challenges identified, this research proposes the use of advanced machine learning algorithms to predict breast cancer outcomes.

02

The study employs the Wisconsin Breast Cancer Dataset (WBCD) as the primary data source and implements various machine learning models, including Logistic Regression, Support Vector Machines (SVM), and Decision Trees.

03

Each model is trained and tested rigorously to assess its performance in predicting whether a breast tumor is benign or malignant. The approach focuses on maximizing accuracy while ensuring the models remain interpretable and practical for clinical use.

LITERATURE REVIEW

LITERATURE REVIEW

The application of machine learning to breast cancer prediction has seen significant progress in recent years.

Various studies have demonstrated the efficacy of algorithms such as SVMs, neural networks, and decision trees in improving diagnostic accuracy. However, there remains a need for models that can be easily integrated into clinical workflows and applied across diverse patient populations.

The review underscores the importance of data preprocessing and feature selection, which are crucial for improving model performance and reliability.

METHODOLOGY

METHODOLOGY

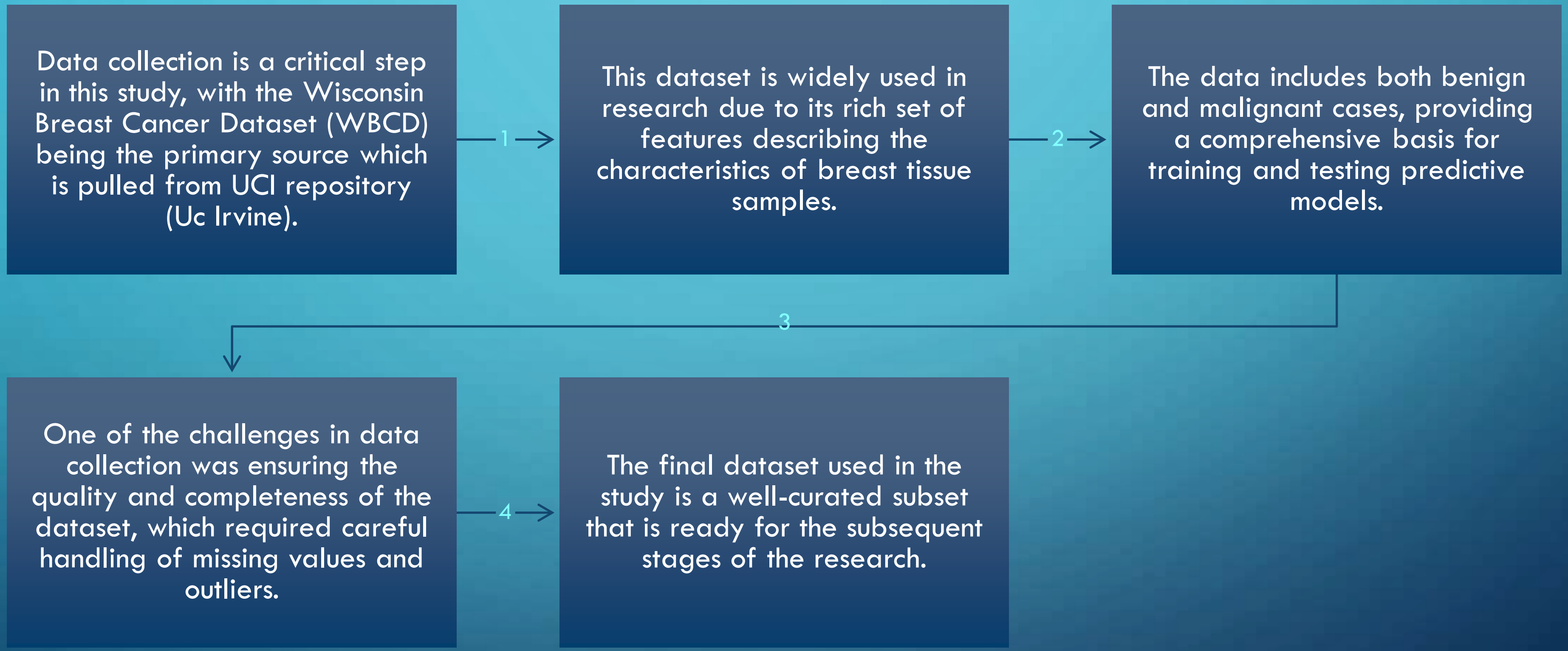
The research adopts the CRISP-DM (Cross-Industry Standard Process for Data Mining) framework, which is widely used for solving business problems through data mining.

The methodology involves several stages, beginning with data understanding and preprocessing, followed by model development and evaluation. The WBCD serves as the primary dataset, and the preprocessing steps include data normalization, handling missing values, and feature selection.

Various machine learning models, including Logistic Regression, SVM, and Decision Trees, are implemented and evaluated using cross-validation and confusion matrix analysis to ensure robustness and accuracy.

DATA COLLECTION

DATA COLLECTION



Data collection is a critical step in this study, with the Wisconsin Breast Cancer Dataset (WBCD) being the primary source which is pulled from UCI repository (Uc Irvine).

This dataset is widely used in research due to its rich set of features describing the characteristics of breast tissue samples.

The data includes both benign and malignant cases, providing a comprehensive basis for training and testing predictive models.

One of the challenges in data collection was ensuring the quality and completeness of the dataset, which required careful handling of missing values and outliers.

The final dataset used in the study is a well-curated subset that is ready for the subsequent stages of the research.

DATA PRE-PROCESSING

DATA PRE-PROCESSING

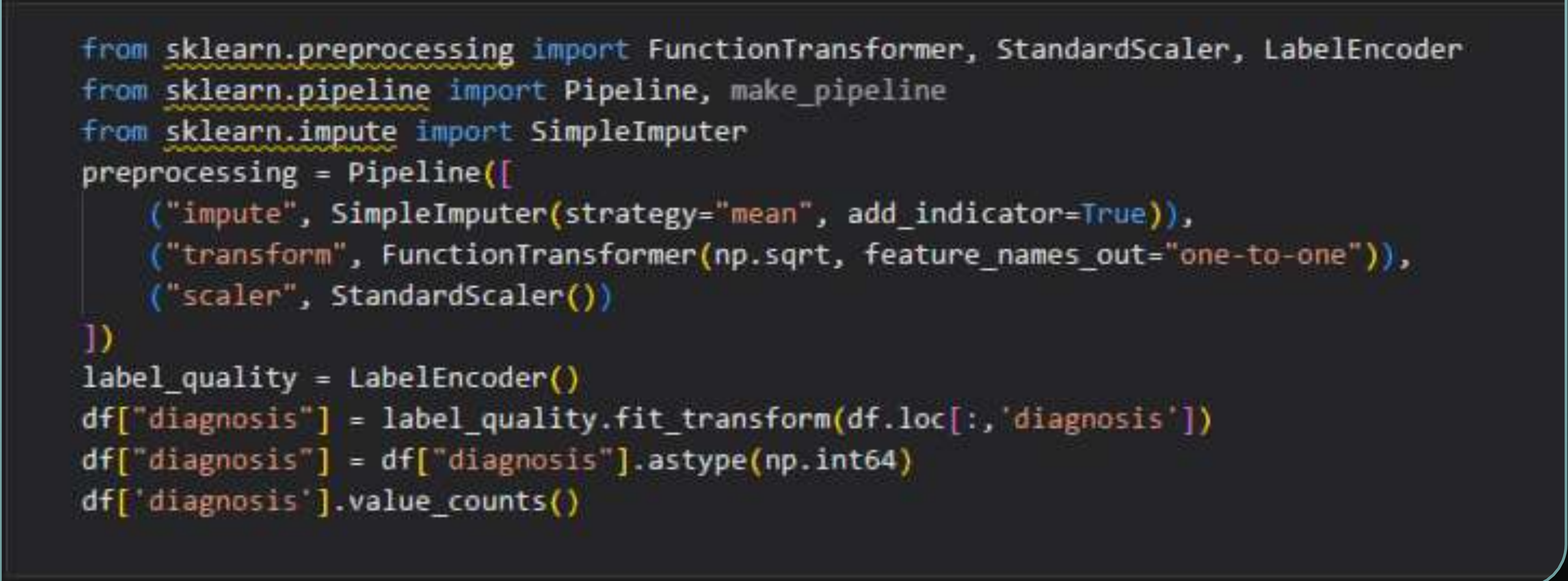
Data preprocessing is a crucial step in preparing the dataset for model training.

This process involves several key activities, including data normalization, which ensures that all features are on a comparable scale.

Additionally, missing data is handled through imputation techniques to ensure that the models have a complete and accurate dataset to work with.

The preprocessing steps are implemented using Python libraries such as Pandas and Scikit-learn, which provide powerful tools for data manipulation and preparation. The outcome of this phase is a clean, well-structured dataset that is ready for modeling.

Preprocessing



```python
from sklearn.preprocessing import FunctionTransformer, StandardScaler, LabelEncoder
from sklearn.pipeline import Pipeline, make_pipeline
from sklearn.impute import SimpleImputer
preprocessing = Pipeline([
    ("impute", SimpleImputer(strategy="mean", add_indicator=True)),
    ("transform", FunctionTransformer(np.sqrt, feature_names_out="one-to-one")),
    ("scaler", StandardScaler())
])
label_quality = LabelEncoder()
df["diagnosis"] = label_quality.fit_transform(df.loc[:,'diagnosis'])
df["diagnosis"] = df["diagnosis"].astype(np.int64)
df['diagnosis'].value_counts()
```

DATA MODELLING

DATA MODELING

The study employs several machine learning models to predict breast cancer outcomes, including Logistic Regression, SVM, and Decision Trees.

Each model is trained on the preprocessed dataset and evaluated using performance metrics such as accuracy, precision, recall, and F1-score.

Logistic Regression is chosen for its simplicity and interpretability, while SVM is known for its effectiveness in high-dimensional spaces.

Decision Trees are included for their ability to capture non-linear relationships in the data.

Cross-validation is used to assess the generalizability of the models, and confusion matrix analysis provides insights into the classification performance, particularly in distinguishing between benign and malignant cases.

TOOLS AND TECHNOLOGIES

TOOLS AND TECHNOLOGIES

The research utilizes a range of tools and technologies to support the data analysis and modeling processes.

Python is the primary programming language used, with key libraries including Pandas, NumPy, Scikit-learn, Matplotlib, and Seaborn for data manipulation, model development, and visualization.

Google collab provide an interactive environment for conducting and documenting the experiments. Additionally, version control tools such as Git are employed to manage the codebase, ensuring that the research process is reproducible and well-organized.

The computational resources required for training the models are supported by high-performance computing systems, ensuring that the experiments are conducted efficiently.

RESULTS
Learn

RESULTS

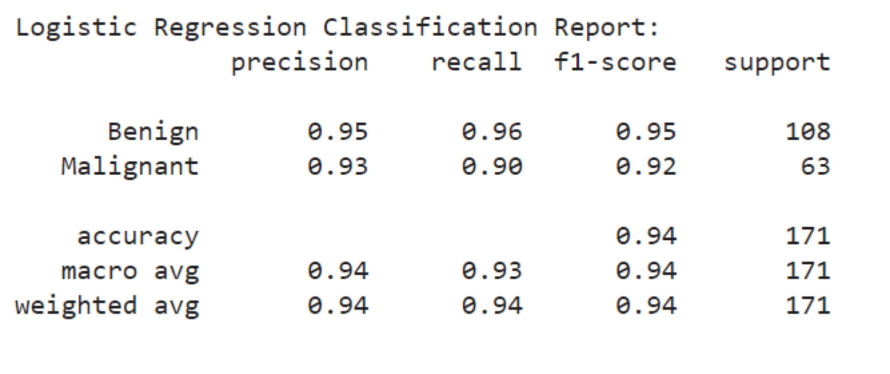
The results of the study demonstrate the effectiveness of the machine learning models in predicting breast cancer outcomes.

Logistic Regression achieved a training accuracy of 94% ,indicating that the model generalizes well to unseen data. The SVM model performed highest number of accuracy, with a training accuracy of 97%.

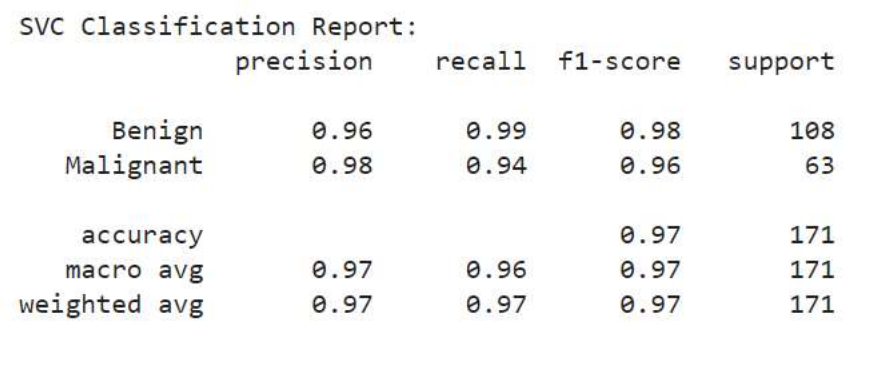
Decision Trees showed training accuracy at 92%, although this may indicate a risk of overfitting. The confusion matrix analysis revealed that the models were able to correctly classify 91% of the cases as benign, with only 9% misclassified as malignant.

These results highlight the potential of machine learning to improve breast cancer prognosis.

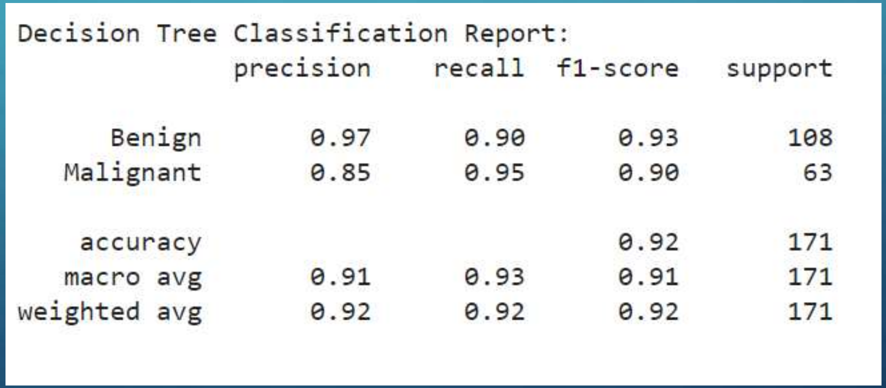
RESULTS



```
Logistic Regression Classification Report:
              precision    recall  f1-score   support

      Benign       0.95      0.96      0.95       108
   Malignant       0.93      0.90      0.92        63

    accuracy                           0.94       171
   macro avg       0.94      0.93      0.94       171
weighted avg       0.94      0.94      0.94       171
```

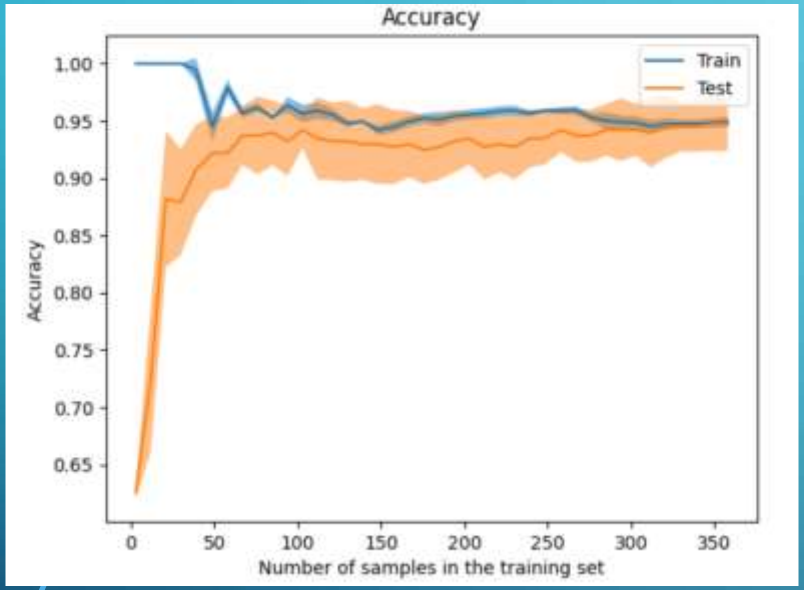


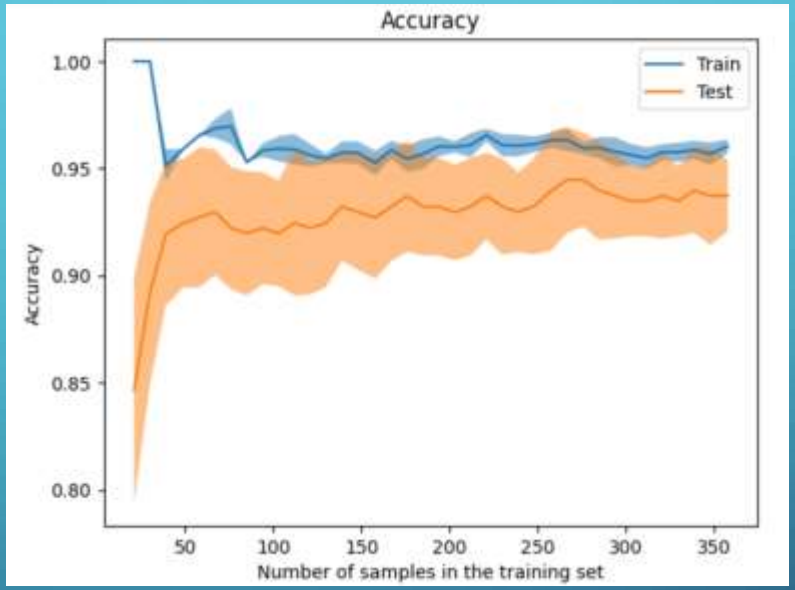
```
SVC Classification Report:
              precision    recall  f1-score   support

      Benign       0.96      0.99      0.98       108
   Malignant       0.98      0.94      0.96        63

    accuracy                           0.97       171
   macro avg       0.97      0.96      0.97       171
weighted avg       0.97      0.97      0.97       171
```

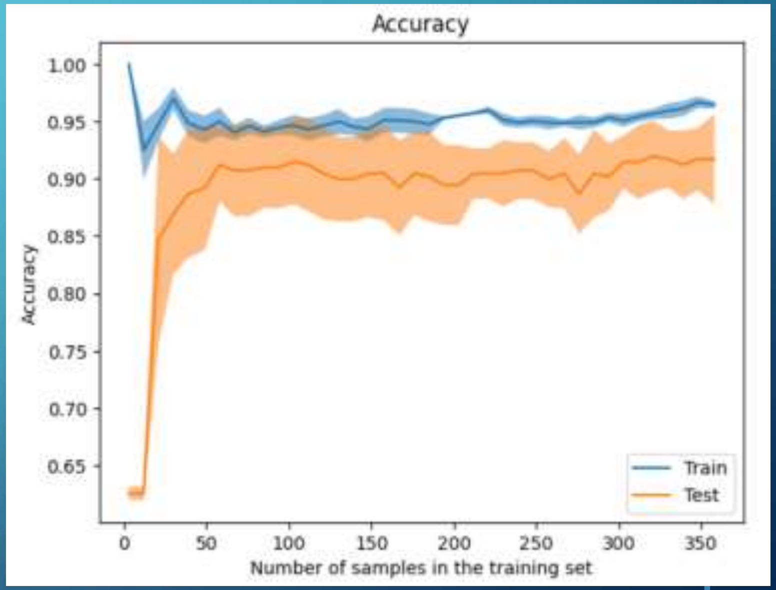


```
Decision Tree Classification Report:
              precision    recall  f1-score   support

      Benign       0.97      0.90      0.93       108
   Malignant       0.85      0.95      0.90        63

    accuracy                           0.92       171
   macro avg       0.91      0.93      0.91       171
weighted avg       0.92      0.92      0.92       171
```

RESULTS







Logistic regression

SVM

Decision Tree

Conclusion

CONCLUSION

- In conclusion, this study demonstrates the potential of machine learning to enhance breast cancer prognosis. By implementing and evaluating several models, the research provides evidence that these techniques can achieve high levels of accuracy, making them valuable tools for clinicians.
- The study's contribution lies in its focus on model interpretability and the rigorous evaluation of performance across multiple metrics. While the results are promising, further research is needed to address the limitations identified and to explore new avenues for improving prediction accuracy.
- The integration of additional data sources and the development of more interpretable models will be key areas of focus for future work.

QUESTIONS