BUSINESS ANALYTICS

CHAPTER – 03

INTRODUCTION TO PRESCRIPTIVE ANALYTICS

Predictive Analytics is the process of using **historical data, statistical algorithms, and machine learning techniques** to predict future outcomes. It helps businesses make **data-driven decisions** by forecasting trends, customer behaviors, and potential risks.

Key Features of Predictive Analytics:

- Uses **past data** to predict future trends.
- Relies on **statistical models and machine learning algorithms**.
- Helps businesses **reduce risks and optimize operations**.

Business Applications:

- **Retail**: Forecasting customer demand to optimize inventory.
- **Finance**: Credit risk assessment and fraud detection.
- **Healthcare**: Predicting disease outbreaks or patient readmissions.
- **Marketing**: Customer churn prediction and personalized recommendations.

Logic-Based and Data-Driven Models in Predictive Analytics

Predictive analytics is a powerful tool that enables businesses to forecast trends, optimize processes, and make informed decisions. At its core, predictive analytics relies on **Logic-Based Models** and **Data-Driven Models** to analyze historical data and predict future outcomes. Both approaches play a crucial role in different business scenarios, and understanding their differences helps organizations choose the best model for their needs.

---

1. Introduction to Predictive Analytics Models

Predictive analytics models aim to extract patterns from historical data to predict future trends and behaviors. These models can be classified into two broad categories:

1. **Logic-Based Models** – Rule-based systems that operate on predefined logic.
2. **Data-Driven Models** – Machine learning-based systems that rely on statistical patterns and data trends.

Each model has its strengths and is suited for specific business applications.

---

2. Logic-Based Models

Definition:

Logic-based models rely on predefined rules and expert knowledge to make decisions. These models use **IF-THEN** rules to determine outcomes based on given conditions.

Characteristics of Logic-Based Models:

✓ Based on human expertise and domain knowledge.
✓ Rules are explicitly programmed by subject matter experts.
✓ Transparent and easy to interpret.
✓ Best suited for problems where conditions and relationships are well understood.

Types of Logic-Based Models:

1. Rule-Based Decision Systems

- Uses **IF-THEN** rules to process information.
- Example: In credit scoring, a bank may define rules such as:
    - **IF** credit score > 750 **AND** income > $50,000, **THEN** approve loan.
    - **IF** credit score < 600, **THEN** reject loan.

2. Expert Systems

- Uses **knowledge bases** to provide recommendations.
- Example: A medical diagnostic system using predefined rules to recommend treatment based on symptoms.

3. Decision Trees (Basic Rule-Based Approach)

- A graphical representation of rules for decision-making.
- Example: A bank determines whether to grant a **home loan** based on income, debt, and employment history.

Advantages of Logic-Based Models:

✓ Easy to interpret and explain to stakeholders.
✓ Works well in structured environments where rules are clear.
✓ Suitable for regulatory compliance and legal applications.

Limitations of Logic-Based Models:

✗ Limited in handling **complex data relationships**.
✗ Cannot adapt to new patterns **without manual intervention**.
✗ Does not perform well with **large, unstructured datasets**.

---

3. Data-Driven Models

Definition:

Data-driven models use **statistical algorithms and machine learning techniques** to uncover patterns in historical data and make predictions. Unlike logic-based models, they **learn from data** without requiring explicit rules.

Characteristics of Data-Driven Models:

✓ Automatically identifies patterns and relationships in data.
✓ Adapts to changing trends without manual rule updates.
✓ Can process large amounts of **unstructured data**.
✓ Uses probability and statistics for decision-making.

Types of Data-Driven Models:

1. Regression Models

- Used for predicting **continuous numerical values**.
- Example: A retailer predicts **future sales** based on advertising spend.
- **Common Algorithms:**
    - **Linear Regression** – Predicts a dependent variable based on one or more independent variables.
    - **Logistic Regression** – Used for binary classification problems (e.g., predicting customer churn).

2. Machine Learning Models

- Uses **algorithms** to train models from data.
- Example: A bank predicts **loan default risk** using customer demographics and credit history.
- **Common Algorithms:**
    - **Decision Trees & Random Forests** – For classification and regression tasks.
    - **Neural Networks & Deep Learning** – For complex pattern recognition.

3. Clustering Models

- Groups similar data points together.
- Example: A telecom company segments customers based on usage patterns.
- **Common Algorithms:**
    - **K-Means Clustering**
    - **Hierarchical Clustering**

Advantages of Data-Driven Models:

✓ Can handle **complex, large-scale data**.
✓ Improves accuracy over time by learning from new data.
✓ Removes human biases in decision-making.

Limitations of Data-Driven Models:

✗ Requires **large datasets** to function effectively.
✗ Often **lacks interpretability**, making it harder to explain decisions.
✗ Computationally intensive and requires **advanced expertise**.

---

4. Key Differences Between Logic-Based and Data-Driven Models

Feature	Logic-Based Models	Data-Driven Models
Basis	Expert-defined rules	Machine learning and statistics
Adaptability	Static, requires manual updates	Dynamic, learns from data
Interpretability	Easy to explain and transparent	Difficult to interpret (black box models)
Complexity	Simple, rule-based	Handles complex relationships
Data Dependency	Works without large data	Requires large datasets
Examples	Rule-based fraud detection, loan approval	Spam detection, customer segmentation

---

5. Business Case Example: Fraud Detection

Scenario:

A bank wants to identify fraudulent transactions.

Logic-Based Approach:

- IF a transaction is **above $10,000** AND the location is **outside the customer's home country**, THEN flag it as **suspicious**.
- IF a transaction occurs **twice in one minute**, THEN block the card.

Limitations:

- Fraudsters can **bypass static rules** by making smaller transactions.
- Rules **do not adapt** to emerging fraud patterns.

Data-Driven Approach:

- The bank uses **machine learning models** trained on past fraudulent transactions.
- The model detects fraud based on **customer behavior, transaction frequency, and device location**.
- If an **unusual spending pattern** is detected, the system **automatically flags** the transaction.

Benefits:
✓ The model learns from new fraud attempts.
✓ Captures **complex patterns** that rule-based systems may miss.

---

6. Choosing the Right Model for Business

Choosing between **logic-based and data-driven models** depends on business needs:

Business Requirement	Best Model Choice
Regulatory Compliance	**Logic-Based Models** (e.g., financial auditing)
Handling Big Data	**Data-Driven Models** (e.g., customer analytics)
Interpretability is Critical	**Logic-Based Models**
High Accuracy Required	**Data-Driven Models** (e.g., predictive maintenance)

---

7. Future of Predictive Analytics Models

🚀 **Hybrid Approaches:** Combining logic-based and data-driven methods for optimal results.
🚀 **AI and Deep Learning:** Improved model performance using neural networks.
🚀 **Explainable AI (XAI):** Making machine learning models more interpretable.
🚀 **Automated Machine Learning (AutoML):** Reducing manual efforts in model training.

PREDICTIVE ANALYTICS: MODELS AND PROCEDURE

Predictive analytics is a branch of advanced analytics that leverages statistical techniques, machine learning, and data mining to predict future trends and behaviors. Businesses use predictive models to **identify patterns**, **forecast outcomes**, and **enhance decision-making**.

---

1. Introduction to Predictive Analytics

Predictive analytics involves analyzing historical data to make future predictions. It helps organizations **mitigate risks**, **optimize operations**, and **enhance customer experiences**.

Key Objectives:

✓ Forecast future trends and outcomes.
✓ Improve decision-making with data-driven insights.
✓ Optimize business strategies for competitive advantage.
✓ Reduce uncertainty by identifying patterns and risks.

Applications in Business:

✓ Fraud detection in banking.
✓ Demand forecasting in retail.
✓ Predictive maintenance in manufacturing.
✓ Customer churn prediction in telecom.

---

2. Predictive Analytics Models

Predictive analytics models use **statistical methods** and **machine learning algorithms** to generate forecasts. These models fall into different categories based on their purpose and complexity.

1. Regression Models

- Used for predicting continuous numerical values.
- Example: **Predicting sales revenue** based on marketing spend.
- **Common Types:**
  - **Linear Regression** – Establishes relationships between variables.
  - **Multiple Regression** – Uses multiple factors to improve accuracy.

2. Classification Models

- Categorizes data into predefined groups.
- Example: **Predicting loan default (Yes/No)** based on credit history.
- **Common Algorithms:**
  - **Logistic Regression** – Predicts probabilities for binary classification.
  - **Decision Trees & Random Forest** – Splits data into logical groups for decision-making.

3. Time Series Models

- Analyzes trends over time to make future predictions.
- Example: **Forecasting stock prices or energy consumption.**

- **Common Techniques:**
    - **ARIMA (Auto-Regressive Integrated Moving Average)**
    - **Exponential Smoothing Models**

4. Clustering Models

- Groups similar data points together for insights.
- Example: **Customer segmentation for targeted marketing.**
- **Common Algorithms:**
    - **K-Means Clustering**
    - **Hierarchical Clustering**

5. Deep Learning Models

- Uses artificial neural networks to detect complex patterns.
- Example: **Image recognition for medical diagnosis.**
- **Common Algorithms:**
    - **Convolutional Neural Networks (CNNs)** – Image and video analysis.
    - **Recurrent Neural Networks (RNNs)** – Sequential data prediction.

---

3. Predictive Analytics Procedure

The predictive analytics process follows a structured **workflow** to ensure accuracy and reliability in forecasting.

Step 1: Define the Business Problem

- Identify **what needs to be predicted** and why.
- Example: A retail company wants to predict **next quarter's sales** to optimize inventory.

Step 2: Data Collection

- Gather **relevant historical data** from multiple sources.
- Data can come from **databases, CRM systems, IoT sensors,** and **third-party APIs**.

Step 3: Data Cleaning and Preprocessing

- Remove **missing values**, **duplicate entries**, and **outliers** to ensure data quality.
- Convert categorical variables into numerical representations.

Step 4: Exploratory Data Analysis (EDA)

- Identify **patterns, trends, and correlations** in the dataset.
- Use **visualization techniques** (charts, histograms, box plots) for insights.

Step 5: Feature Selection & Engineering

- Choose the most **relevant variables** for the model.
- Create **new features** that improve predictive power.

Step 6: Select the Predictive Model

- Choose an appropriate model based on the **problem type (regression, classification, etc.)**.
- Example: **Logistic regression** for **churn prediction**, **time series analysis** for **demand forecasting**.

Step 7: Train the Model

- Split data into **training and testing sets** (e.g., 80-20 split).
- Train the model on historical data and fine-tune parameters.

Step 8: Model Evaluation and Validation

- Use **accuracy metrics** (e.g., RMSE, R^2, Precision, Recall) to assess model performance.
- Perform **cross-validation** to prevent overfitting.

Step 9: Model Deployment

- Deploy the trained model in a **live business environment**.
- Example: Integrating a fraud detection model into **banking software**.

Step 10: Monitor and Improve Model Performance

- Continuously monitor predictions and **retrain the model** with new data.
- Adjust parameters if accuracy declines over time.

---

4. Business Case Example: Customer Churn Prediction

Scenario:

A telecom company wants to **predict customer churn** (customers who might leave).

Predictive Analytics Procedure:

✅ **Step 1: Define the Business Problem**

- Predict which customers are likely to **cancel their subscription**.

✅ **Step 2: Data Collection**

- Customer data: **Usage history, call duration, complaints, billing records**.

✅ **Step 3: Data Cleaning & Preprocessing**

- Handle **missing values** in call logs.

✅ **Step 4: Exploratory Data Analysis**

- Find **patterns** (e.g., most churners had **poor customer service ratings**).

✅ **Step 5: Feature Selection & Engineering**

- Use features like **billing issues, internet speed, and call drops**.

✅ **Step 6: Model Selection**

- Use **Logistic Regression** for churn classification.

✅ **Step 7: Train the Model**

- Split data into **80% training, 20% testing**.

✅ **Step 8: Model Evaluation**

- Accuracy: **85%**, Precision: **0.89**, Recall: **0.82**.

✅ **Step 9: Deployment**

- Implement model into **customer support software** for early intervention.

✅ **Step 10: Monitor and Improve**

- Re-train model every **3 months** with updated customer data.

---

5. Advantages of Predictive Analytics

✓ **Improves Decision-Making:** Helps businesses make data-driven choices.
✓ **Enhances Efficiency:** Automates forecasting processes.
✓ **Reduces Risks:** Identifies potential threats and frauds.
✓ **Increases Profitability:** Optimizes operations for better revenue.

DATA MINING FOR PREDICTIVE ANALYTICS

1. Introduction to Data Mining

Data mining is the process of extracting **useful patterns, trends, and insights** from large datasets. It plays a crucial role in **predictive analytics** by identifying **hidden relationships** that help in forecasting future outcomes.

Key Objectives of Data Mining:

✓ Discover patterns in historical data.
✓ Improve decision-making with predictive models.
✓ Enhance business intelligence and strategic planning.
✓ Detect anomalies, correlations, and trends in data.

---

2. Importance of Data Mining in Predictive Analytics

Predictive analytics relies on **high-quality data** to generate accurate forecasts. Data mining techniques help:

- **Preprocess raw data** (cleaning, transformation).
- **Identify key variables** affecting outcomes.
- **Reduce noise** and improve model performance.
- **Optimize machine learning algorithms** by providing refined data.

---

3. Data Mining Techniques for Predictive Analytics

1. Classification

- Categorizes data into predefined groups.
- Example: **Predicting customer churn (Will Leave/Will Stay)**.
- **Common Algorithms:**
    - Decision Trees
    - Random Forest
    - Support Vector Machines (SVM)

2. Regression

- Predicts numerical outcomes based on input variables.
- Example: **Forecasting sales revenue for next quarter**.
- **Common Techniques:**
    - Linear Regression
    - Multiple Regression
    - Ridge & Lasso Regression

3. Clustering

- Groups similar data points together based on common characteristics.
- Example: **Customer segmentation for targeted marketing.**
- **Common Algorithms:**

- K-Means Clustering
- Hierarchical Clustering

4. Association Rule Mining

- Identifies relationships between variables.
- Example: **Market Basket Analysis (Customers who buy bread also buy butter).**
- **Common Techniques:**
    - Apriori Algorithm
    - FP-Growth Algorithm

5. Anomaly Detection

- Detects **outliers** or unusual patterns in data.
- Example: **Fraud detection in banking transactions.**
- **Common Algorithms:**
    - Isolation Forest
    - Local Outlier Factor

6. Neural Networks & Deep Learning

- Advanced machine learning techniques for complex pattern recognition.
- Example: **Speech and image recognition.**
- **Common Models:**
    - Convolutional Neural Networks (CNNs)
    - Recurrent Neural Networks (RNNs)

---

4. Data Mining Process for Predictive Analytics

Step 1: Business Understanding

- Define the problem and objectives.
- Example: A telecom company wants to predict **customer churn**.

Step 2: Data Collection

- Gather relevant data from multiple sources (databases, APIs, sensors).
- Example: **Customer complaints, usage history, service feedback**.

Step 3: Data Preprocessing

- Remove missing values, clean duplicates, and normalize data.

- Convert categorical data into numerical formats.

Step 4: Data Exploration & Feature Selection

- Perform **Exploratory Data Analysis (EDA)** to find **important patterns**.
- Select the most relevant features for predictive modeling.

Step 5: Model Selection & Training

- Choose the best data mining model based on the problem.
- Example: **Logistic Regression for churn prediction**.

Step 6: Model Evaluation & Validation

- Use metrics like **Accuracy, Precision, Recall, and F1-score** to measure performance.
- Perform **cross-validation** to ensure reliability.

Step 7: Deployment & Continuous Improvement

- Deploy the model into a live system for real-time predictions.
- Regularly retrain the model with **new data**.

---

5. Business Case Example: Fraud Detection in Banking

Scenario:

A bank wants to detect fraudulent transactions in real time.

Data Mining Process Applied:

✅ **Step 1: Business Understanding**

- The goal is to identify **suspicious transactions** automatically.

✅ **Step 2: Data Collection**

- Transaction history, user behavior, location data.

✅ **Step 3: Data Preprocessing**

- Normalize transaction values and remove **duplicate transactions**.

✅ **Step 4: Data Exploration & Feature Engineering**

- Identify **patterns** in fraudulent transactions.

✅ **Step 5: Model Selection**

- Use **Anomaly Detection (Isolation Forest Algorithm)**.

✅ **Step 6: Model Evaluation**

- Measure **False Positives & True Positives**.

✅ **Step 7: Deployment**

- Integrate fraud detection into the **bank's security system**.

✅ **Step 8: Continuous Monitoring**

- Improve the model based on new fraud trends.

---

6. Advantages of Data Mining in Predictive Analytics

✓ **Enhances Decision-Making:** Provides data-driven insights.
✓ **Reduces Operational Risks:** Detects fraud, inefficiencies, and anomalies.
✓ **Improves Customer Experience:** Enables personalized recommendations.
✓ **Optimizes Business Processes:** Automates pattern recognition.

REGRESSION ANALYSIS: STUDENT PERFORMANCE VS. SLEEPING HOURS

1. Introduction to Regression Analysis

Regression analysis is a statistical technique used to examine the relationship between **dependent and independent variables**. In this case, we analyze how **sleeping hours impact student performance in an exam** using **linear regression**.

Key Components:

- **Dependent Variable (Y):** Student's Exam Score
- **Independent Variable (X):** Number of Sleeping Hours
- **Regression Equation:** $Y=\beta_0+\beta_1 X+\epsilon$

  where:

  - β0\beta_0β0 = Intercept
  - β1\beta_1β1 = Coefficient for Sleeping Hours
  - ϵ\epsilonϵ = Error term

---

2. Business Case: Predicting Exam Scores

Scenario:

A university wants to determine the impact of **sleep duration** on **student performance**. The goal is to use regression analysis to find a trend between **sleeping hours** and **exam scores**, which can help improve student learning strategies.

Hypothesis Testing:

- **Null Hypothesis (H_0):** Sleeping hours do not affect exam performance.

- **Alternative Hypothesis (H_1):** Sleeping hours significantly impact exam performance.

---

3. Regression Analysis: Steps

Step 1: Data Collection

The university collects data from 100 students on:

- Average **sleeping hours per day** during exam preparation.
- **Exam scores** (out of 100).

Step 2: Data Visualization (Scatter Plot)

A scatter plot is used to visualize the relationship between **sleeping hours** and **exam scores**. If there is a trend (e.g., positive or negative correlation), regression analysis can be applied.

Step 3: Regression Model Construction

Using a **simple linear regression model**, we fit the equation:

$\text{Exam Score} = \beta_0 + \beta_1(\text{Sleeping Hours}) + \epsilon$

Example Output:

- $\beta_0 = 40$ (Intercept)
- $\beta_1 = 5$ (Coefficient)

Regression Equation:

$\text{Exam Score} = 40 + 5(\text{Sleeping Hours})$Exam

This means that **for every additional hour of sleep, the expected exam score increases by 5 points**.

---

4. Regression Statistics & ANOVA Analysis

Regression Statistics Table

Statistic	Value
Multiple R	0.75
R^2 (R-Squared)	0.56
Adjusted R^2	0.55
Standard Error	5.2
Observations	100

- **R² Value:** 0.56 → Indicates that **56% of the variation** in exam scores is explained by sleeping hours.
- **Multiple R:** 0.75 → Shows a **strong positive correlation**.

ANOVA (Analysis of Variance) Table

Source	SS (Sum of Squares)	df (Degree of Freedom)	MS (Mean Square)	F-Statistic	p-value
Regression	3400	1	3400	42.5	0.0001
Residual	2700	98	27.5	-	-
Total	6100	99	-	-	-

- **F-Statistic (42.5) & p-value (0.0001):** Since the p-value is **less than 0.05**, the model is **statistically significant** → Sleep duration **affects** exam scores.

---

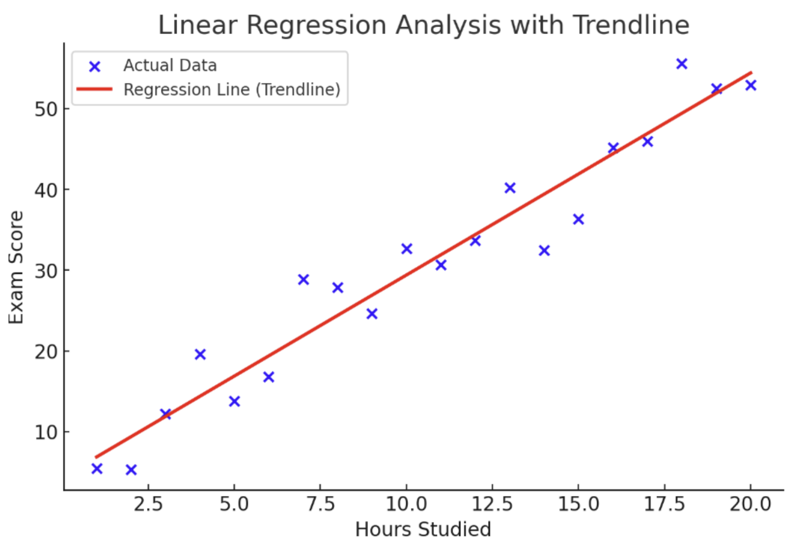
5. Interpretation of Results

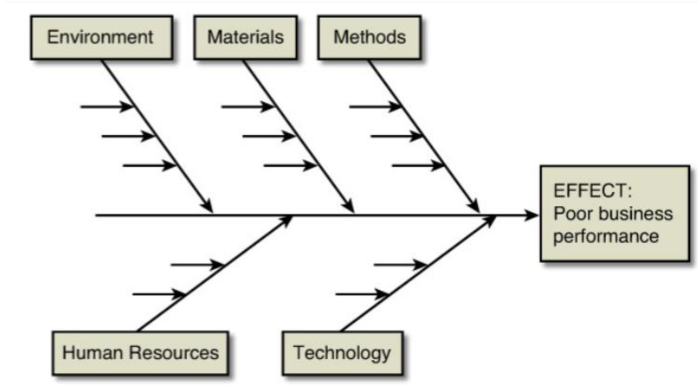
- **Positive Correlation:** More sleep hours **lead to better exam scores**.
- **Significance:** Since p-value < 0.05, sleep **significantly influences** student performance.
- **Practical Implication:** Universities can use these insights to **recommend optimal sleep durations** for students before exams.

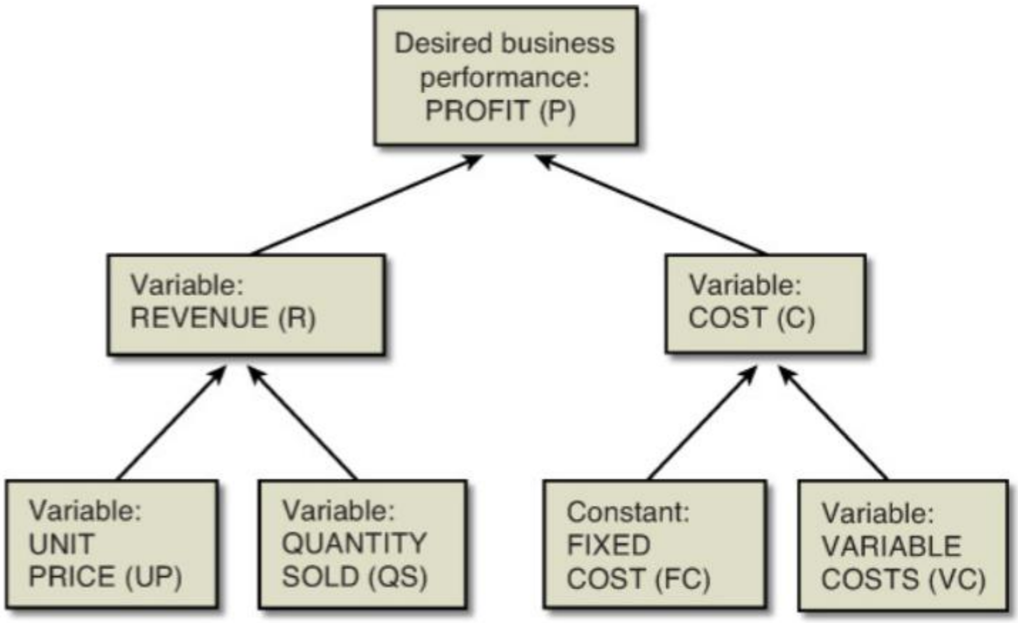
---

6. Conclusion

This regression analysis confirms that **sleeping hours have a direct impact on exam performance**. By optimizing sleep schedules, students can **improve concentration, retention, and scores**.




Environment
Materials
Methods
EFFECT:
Poor business
performance
Human Resources
Technology


Desired business
performance:
PROFIT (P)
Variable:
REVENUE (R)
Variable:
COST (C)
Variable:
UNIT
PRICE (UP)
Variable:
QUANTITY
SOLD (QS)
Constant:
FIXED
COST (FC)
Variable:
VARIABLE
COSTS (VC)