ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

Московский институт электроники и математики им. А.Н. Тихонова

Шембель Даниил Альбертович, группа БИТ213

Модульная работа №2

Семестр 2

по дисциплине «Прикладной Статистический анализ данных»

Дата сдачи отчета: 23.05.23

Москва 2023 г.

Оглавление

Актуальность

Изучение состояния экономики мировых держав позволяет сравнивать экономическое развитие различных стран и выявлять ключевые факторы, влияющие на успешность экономики. Результаты исследования могут быть использованы для прогнозирования и планирования экономического развития, а также для принятия обоснованных решений в области политики и инвестиций. Оно также помогает понять причины неравенства в экономическом развитии между странами и способствует разработке стратегий для сокращения этого разрыва. В целом, исследование имеет значимость для мировой экономики и принятия обоснованных решений.

Исследуемые показатели

1. Индекс потребительских цен, 2010 г. = 100: Мера изменения цен на потребительские товары и услуги с базовым уровнем в 2010 году.
2. Экспорт товаров и услуг, доля в процентах от ВВП: Процентный вклад экспорта в общую стоимость производства и услуг в стране.
3. Импорт товаров и услуг, доля в процентах от ВВП: Процентный объем импорта по отношению к общей стоимости производства и услуг в стране.
4. Обменный курс, единиц национальной валюты за доллар США: Отношение стоимости национальной валюты к доллару США, определяющее покупательную способность и международную конкурентоспособность страны.
5. Уровень безработицы, %: Процент людей без работы в экономике страны.
6. ВВП на душу населения в ценах и ППС (США) 2010 г., темп роста: Средняя стоимость производства на одного человека, учитывающая разницу в уровне цен и покупательной способности с учетом уровня цен в США, и ее изменение со временем.
7. Внешний баланс товаров и услуг (бинарная переменная 1-положительный, 0-отрицательный): Измерение разницы между стоимостью экспорта и импорта товаров и услуг, где 1 указывает на положительный баланс, а 0 на отрицательный баланс.
8. Уровень экономического развития страны (задача классификации) (ИРЧП > 0,85 = 1, < 0,85 = 0): Классификация страны на основе ее уровня экономического развития, где значение 1 означает высокий уровень развития, а значение 0 - низкий уровень развития.

Вступление

В современном мире экономическое развитие стран является одним из главных приоритетов для правительств, международных организаций и инвесторов. Понимание факторов, влияющих на успешность экономики, и способность прогнозировать ее развитие имеют огромное значение для принятия обоснованных решений в области политики, инвестиций и развития. В этом контексте проведение исследований, направленных на оценку экономического развития стран, становится особенно актуальным.

В рамках таких исследований использование статистических методов, таких как дискриминантный анализ и построение дерева решений, позволяет систематически анализировать и классифицировать различные аспекты экономического развития. Такой подход позволяет выявить ключевые факторы, влияющие на успешность экономики, и установить связи между этими факторами и развитием стран.

В контексте описанного исследования были рассмотрены несколько важных признаков, включая индекс потребительских цен, долю экспорта и импорта от ВВП, обменный курс, уровень безработицы,

ВВП на душу населения, внешний баланс товаров и услуг, а также уровень экономического развития. Анализ и классификация этих признаков предоставляют ценную информацию о текущем состоянии экономики и ее потенциале для роста.

Цель данного исследования заключается в понимании того, как эти различные факторы влияют на экономическое развитие стран и как они могут быть использованы для прогнозирования будущих тенденций. Полученные результаты могут послужить основой для разработки эффективных стратегий, ориентированных на улучшение экономической ситуации в различных странах и сокращение разрыва в развитии между ними.

В итоге, проведение дискриминантного анализа и построение дерева решений для оценки экономического развития стран имеет важное значение для мировой экономики, планирования и принятия обоснованных решений. Предоставление обширной информации о факторах, влияющих на экономику, позволяет лучше понять ее состояние, выявить потенциал для роста и разработать стратегии, способствующие устойчивому и равномерному развитию стран.

Дискриминантный анализ

Сначала была проведена кластеризация методом k-means и были выделены 3 страны с наибольшим расстоянием до центра своего кластера- Беларусь, ЦАР, Хорватия. Они были исключены из общей выборки. Далее был проведен дискриминантный анализ и были получены коэффициенты дискриминантной функции для каждого параметра:

```
Относительный вклад каждой переменной:
Индекс потребительских цен, 2010 г.=100: 0.636458883978744
Экспорт товаров и услуг, доля в процентах от ВВП: -0.35277360676557407
Импорт товаров и услуг, доля в процентах от ВВП: 0.42539948985529535
Обменный курс, единиц национальной валюты за долл. США: 0.004381226674423885
Уровень безработицы, %: 1.0437378905331687
ВВП на душу населения в ценах и ППС(США) 2010 г., темп роста: -0.7572038842760579
```

А также были получены средние значения дискриминатной функции для классов:

```
Средние значения дискриминантной функции по группам:
Группа 0: -0.6613715756598637
Группа 1: 0.785378746096087
```

Далее было получено значение лямбды Уилкса:

```
lambda_wilks = lda.score(features, target)
print(lambda_wilks)
```

```
0.7428571428571429
```

, по итогам данного значения можно сделать следующие выводы:

Существует некоторая мультиколлинеарность между зависимыми переменными. Однако, уровень коллинеарности не является существенным, и данные все же содержат значимую информацию для анализа. Возможно, некоторые из зависимых переменных взаимосвязаны друг с другом, что может затруднить точное определение их влияния на целевую переменную. Для дальнейшего анализа и интерпретации результатов дискриминантного анализа рекомендуется учитывать присутствие мультиколлинеарности и быть осторожными при делении значимости между зависимыми переменными.

В целом, значение лямбда Уилкса 0.74 указывает на наличие мультиколлинеарности, но не до такой степени, чтобы полностью исключать зависимые переменные из анализа.

Далее была проведена попытка предсказать класс для 3 стран, ранее исключенных из исследования:

	Страна	Реальный класс	Предсказанный класс	Вероятности
0	Беларусь	0	0	[0.9999999962916455, 3.7083545256441943e-09]
1	ЦАР	0	0	[1.0, 1.8658705788479776e-28]
2	Хорватия	0	1	[0.39180122492122915, 0.6081987750787708]

Видно, что мы смогли предсказать класс верно для двух из трех стран: Беларуси и ЦАР с большой долей вероятности (близка к одному, но вследствие ограничения отображения просто 1.0), что несомненно является отличным результатом, однако для Хорватии класс был определен неточно, еще и с большей долей вероятности она была определена в неверный класс, что даёт понять, что модель еще далека от идеала, а также, возможно указывает на недостаток исходных данных.

Выводы

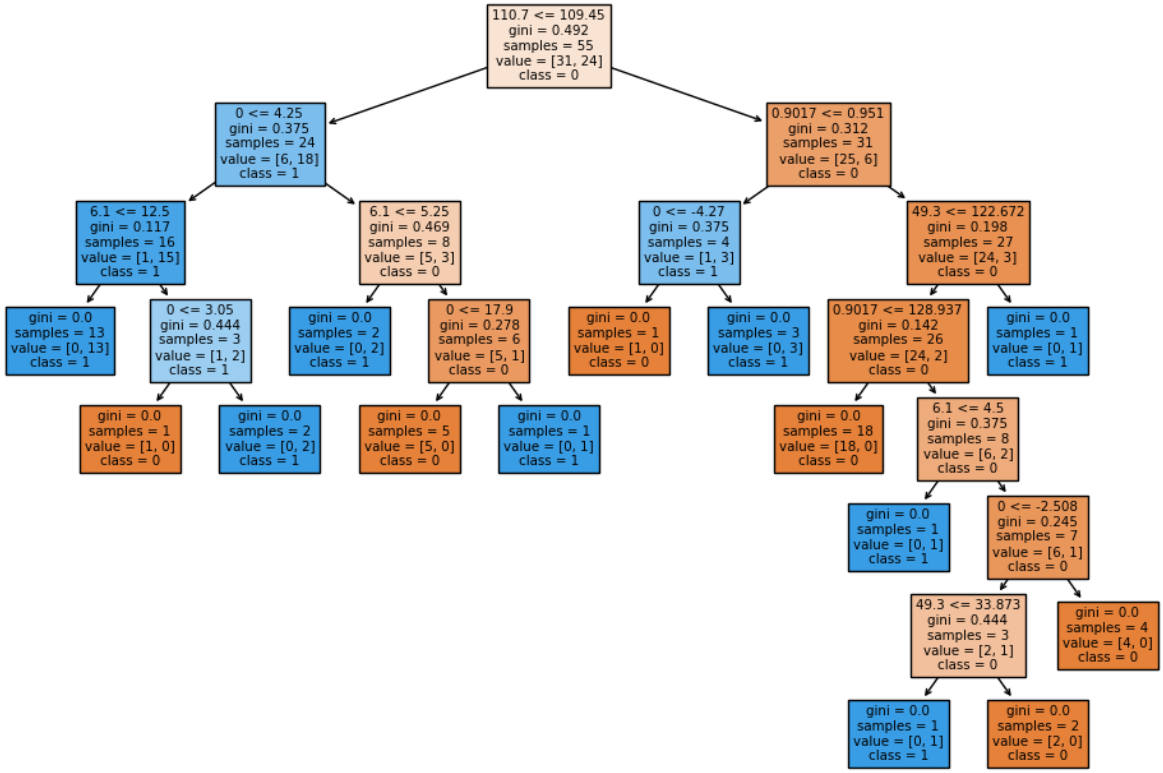
Исходя из относительного вклада каждой переменной в проведенном дискриминантном анализе, можно сделать следующие выводы:

- Индекс потребительских цен, 2010 г.=100 имеет достаточно значительный положительный вклад (0.636458883978744) в разделение групп. Это может указывать на то, что различия в индексе потребительских цен могут играть важную роль в определении принадлежности объектов к разным группам.
- Экспорт товаров и услуг, доля в процентах от ВВП имеет небольшой отрицательный вклад (-0.35277360676557407). Это может означать, что различия в доле экспорта товаров и услуг относительно ВВП не являются существенными факторами в разделении групп.
- Импорт товаров и услуг, доля в процентах от ВВП имеет положительный вклад (0.42539948985529535), что может указывать на то, что различия в доле импорта товаров и услуг могут играть роль в определении принадлежности объектов к разным группам.
- Обменный курс, единиц национальной валюты за доллар США имеет незначительный положительный вклад (0.004381226674423885), что указывает на то, что различия в обменных курсах между национальной валютой и долларом США могут иметь ограниченно слабое значение в разделении групп.
- Уровень безработицы, % имеет существенный положительный вклад (1.0437378905331687), что указывает на то, что различия в уровне безработицы могут играть важную роль в определении принадлежности объектов к разным группам.
- ВВП на душу населения в ценах и ППС (США) 2010 г., темп роста имеет существенный отрицательный вклад (-0.7572038842760579), что может указывать на то, что различия в темпах роста ВВП на душу населения могут играть важную роль в разделении групп.

В целом, проведенный дискриминантный анализ позволяет увидеть, какие переменные имеют наибольший относительный вклад в разделение групп. Это помогает понять, какие факторы могут быть наиболее значимыми при классификации объектов и может быть полезно для дальнейшего исследования и принятия решений.

Дерево решений

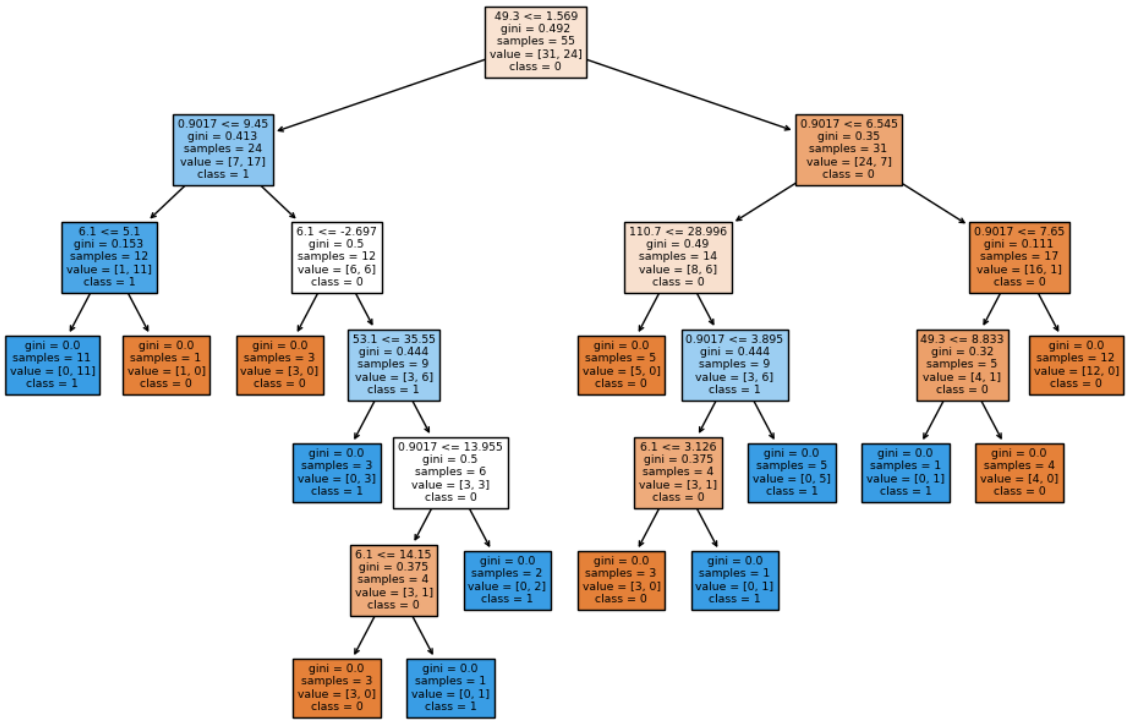
Далее было построение CRT дерево решений. Изначальная выборка была разделена на 2 подвыборки, основная и тестовая (тестовая составляла ≈ 20% от изначальной и выбиралась случайным образом каждый раз). Полученное дерево на основе всех параметров:



А также прогон тестовой выборки по построенному дереву:

```
Accuracy: 0.7857142857142857
Country: Литва - True Class: 0 - Predicted Class: 1
Country: Азербайджан - True Class: 0 - Predicted Class: 1
Country: Швейцария - True Class: 1 - Predicted Class: 1
Country: Бельгия - True Class: 1 - Predicted Class: 1
Country: Австралия - True Class: 1 - Predicted Class: 1
Country: Канада - True Class: 1 - Predicted Class: 1
Country: Грузия - True Class: 0 - Predicted Class: 0
Country: Румыния - True Class: 0 - Predicted Class: 0
Country: Финляндия - True Class: 1 - Predicted Class: 1
Country: Израиль - True Class: 1 - Predicted Class: 1
Country: Республика Молдова - True Class: 0 - Predicted Class: 1
Country: Греция - True Class: 1 - Predicted Class: 1
Country: Ирак - True Class: 0 - Predicted Class: 0
Country: Болгария - True Class: 0 - Predicted Class: 0
```

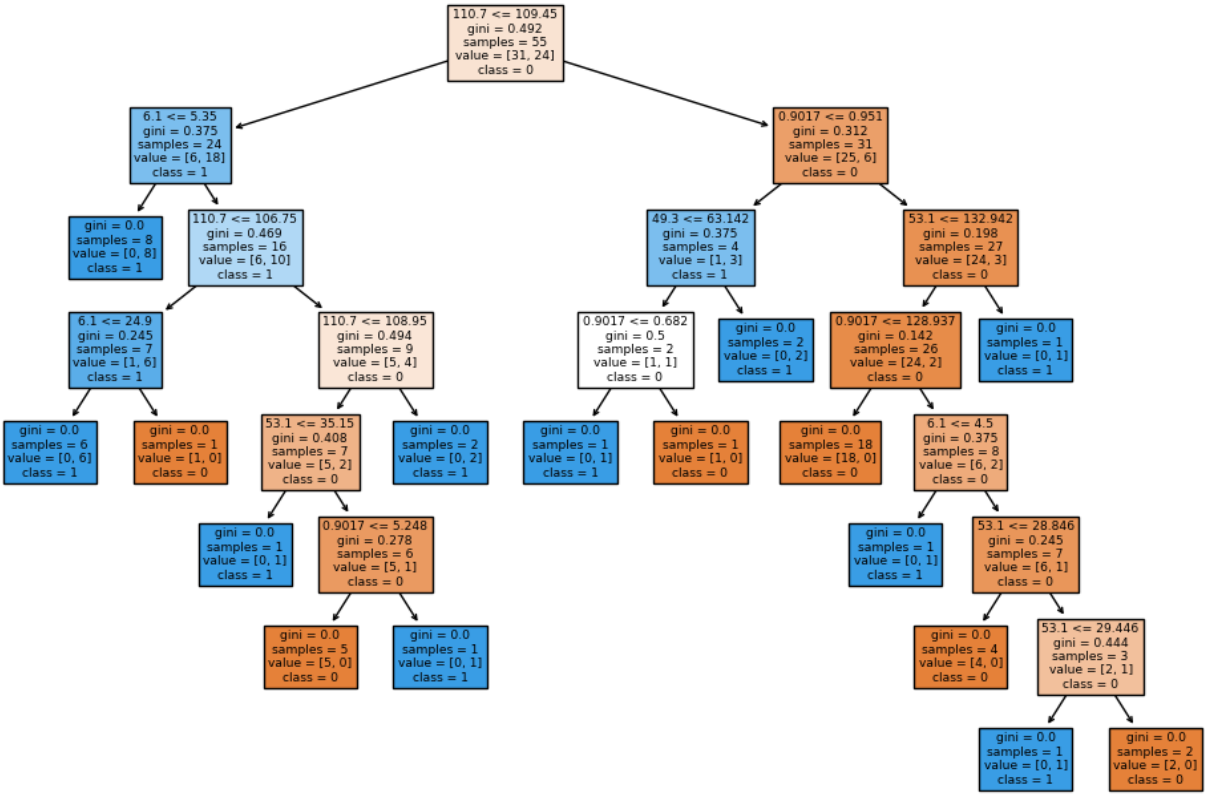
Как можно видеть, модель оказалось точна в ≈78% тестовых случаев. Далее попробуем убрать из параметров Индекс потребительских цен, 2010 г.=100, как один из самых влиятельных факторов (исходя из дискриминантного анализа). Полученное дерево:



А также прогон тестовой выборки по построенному дереву:

```
Accuracy: 0.7857142857142857
Country: Литва - True Class: 0 - Predicted Class: 1
Country: Азербайджан - True Class: 0 - Predicted Class: 1
Country: Швейцария - True Class: 1 - Predicted Class: 1
Country: Бельгия - True Class: 1 - Predicted Class: 1
Country: Австралия - True Class: 1 - Predicted Class: 1
Country: Канада - True Class: 1 - Predicted Class: 1
Country: Грузия - True Class: 0 - Predicted Class: 0
Country: Румыния - True Class: 0 - Predicted Class: 0
Country: Финляндия - True Class: 1 - Predicted Class: 1
Country: Израиль - True Class: 1 - Predicted Class: 1
Country: Республика Молдова - True Class: 0 - Predicted Class: 1
Country: Греция - True Class: 1 - Predicted Class: 1
Country: Ирак - True Class: 0 - Predicted Class: 0
Country: Болгария - True Class: 0 - Predicted Class: 0
```

Как видим, точность оказалась такой же, как и в прошлом пункте, что достаточно странно. В третьем случае уберем ВВП на душу населения в ценах и ППС (США) 2010 г., темп роста, который также оказался существенным параметром по итогам дискриминантного анализа. Полученное дерево:



А также прогон тестовой выборки по построенному дереву:

```
Accuracy: 0.6428571428571429
Country: Литва - True Class: 0 - Predicted Class: 0
Country: Азербайджан - True Class: 0 - Predicted Class: 0
Country: Швейцария - True Class: 1 - Predicted Class: 1
Country: Бельгия - True Class: 1 - Predicted Class: 0
Country: Австралия - True Class: 1 - Predicted Class: 0
Country: Канада - True Class: 1 - Predicted Class: 1
Country: Грузия - True Class: 0 - Predicted Class: 0
Country: Румыния - True Class: 0 - Predicted Class: 0
Country: Финляндия - True Class: 1 - Predicted Class: 0
Country: Израиль - True Class: 1 - Predicted Class: 1
Country: Республика Молдова - True Class: 0 - Predicted Class: 0
Country: Греция - True Class: 1 - Predicted Class: 0
Country: Ирак - True Class: 0 - Predicted Class: 0
Country: Болгария - True Class: 0 - Predicted Class: 1
```

Видим, что итоговая точность упала по сравнению с прошлыми пунктами.

Выводы

Исходя из значения точности (Accuracy) равной 0.7857 на тестовой выборке, можно сделать следующие выводы о построенном дереве решений. Дерево решений достигло достаточно высокой точности прогнозирования на тестовой выборке, примерно 78.57% наблюдений были классифицированы

правильно. Это говорит о том, что модель может достаточно хорошо обобщать данные и прогнозировать классы для новых наблюдений. Однако, стоит отметить, что точность 0.7857 означает, что около 21.43% наблюдений были неправильно классифицированы моделью. Это может указывать на наличие некоторой степени ошибки или шума в данных, а также возможно на необходимость улучшения модели или использования других методов классификации, если требуется более высокая точность.