

*IBM SPSS Modeler 18.3 In-Database* 마  
이닝 안내서



### 참고

이 정보와 이 정보가 지원하는 제품을 사용하기 전에, [85 페이지의 『주의사항』](#)에 있는 정보를 확인하십시오.

### 제품 정보

이 개정판은 새 개정판에 별도로 명시하지 않는 한, IBM® SPSS® Modeler 18, 릴리스 3, 수정사항 0 및 모든 후속 릴리스와 수정에 적용됩니다.

© Copyright International Business Machines Corporation .

# 목차

서론.....	vii
---------	-----

제 1 장 IBM SPSS Modeler 정보 .....	1
---------------------------------	---

IBM SPSS Modeler 제품.....	1
IBM SPSS Modeler .....	1
IBM SPSS Modeler Server .....	1
IBM SPSS Modeler Administration Console .....	1
IBM SPSS Modeler Batch .....	2
IBM SPSS Modeler Solution Publisher .....	2
IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터 .....	2
IBM SPSS Modeler 에디션.....	2
문서.....	3
SPSS Modeler Professional 문서.....	3
SPSS Modeler Premium 문서.....	3
애플리케이션 예제.....	3
Demos 폴더.....	4
라이센스 추적.....	4

제 2 장 In-Database 마이닝.....	5
----------------------------	---

데이터베이스 모델링 개요.....	5
필요사항.....	5
모델 작성.....	6
데이터 준비.....	6
모델 스코어링.....	6
데이터베이스 모델 내보내기 및 저장.....	6
모델 일관성.....	7
생성된 SQL 보기 및 내보내기.....	7

제 3 장 Microsoft Analysis Services를 사용한 데이터베이스 모델링.....	9
--	---

IBM SPSS Modeler 및 Microsoft Analysis Services.....	9
Microsoft Analysis Services와의 통합을 위한 요구사항.....	10
Analysis Services와의 통합 사용.....	11
Analysis Services를 사용하여 모델 작성.....	12
Analysis Services 모델 관리.....	12
모든 알고리즘 노드에 공통인 설정.....	14
MS 의사결정 트리 고급 옵션.....	14
MS 군집화 고급 옵션.....	14
MS Naive Bayes 고급 옵션.....	15
MS 선형 회귀 고급 옵션.....	15
MS 신경망 고급 옵션.....	15
MS 로지스틱 회귀분석 고급 옵션.....	15
MS 연관 규칙 노드.....	15
MS 시계열 노드.....	15
MS 시퀀스 군집화 노드.....	17
Analysis Services 모델 스코어링.....	18
모든 Analysis Services 모델에 공통인 설정.....	18
MS 시계열 모델 너깃.....	18
MS 시퀀스 군집화 모델 너깃.....	19
모델 내보내기 및 노드 생성.....	20
Analysis Services 마이닝 예제.....	20

예제 스트림: 의사결정 트리.....	20
----------------------	----

<b>제 4 장 Oracle Data Mining을 사용한 데이터베이스 모델링.....</b>	<b>23</b>
Oracle Data Mining 정보.....	23
Oracle과의 통합을 위한 요구사항.....	23
Oracle과의 통합 사용.....	24
Oracle Data Mining을 사용하여 모델 작성.....	25
Oracle 모형 서버 옵션.....	26
오분류 비용.....	26
Oracle Naive Bayes.....	26
Naive Bayes 모형 옵션.....	26
Naive Bayes 고급 옵션.....	27
Oracle 적응형 베이스.....	27
적응형 Bayes 모델 옵션.....	27
적응형 Bayes 고급 옵션.....	28
Oracle 지원 벡터 머신(SVM).....	28
Oracle SVM 모형 옵션.....	28
Oracle SVM 고급 옵션.....	29
Oracle SVM 가중치 옵션.....	29
Oracle 일반화 선형 모형(GLM).....	30
Oracle GLM 모형 옵션.....	30
Oracle GLM 고급 옵션.....	30
Oracle GLM 가중치 옵션.....	31
Oracle 의사결정 트리.....	31
의사결정 트리 모형 옵션.....	31
의사결정 트리 고급 옵션.....	31
Oracle O-Cluster.....	32
O-군집 모형 옵션.....	32
O-군집 고급 옵션.....	32
Oracle K-평균.....	32
K-평균 모형 옵션.....	33
K-평균 고급 옵션.....	33
Oracle 비음수 교차표 분해(NMF).....	33
NMF 모형 옵션.....	33
NMF 고급 옵션.....	34
Oracle Apriori.....	34
Apriori 필드 옵션.....	34
Apriori 모형 옵션.....	35
Oracle 최소 설명 길이(MDL).....	35
MDL 모형 옵션.....	36
Oracle 속성 중요도(AI).....	36
AI 모델 옵션.....	36
AI 선택 옵션.....	36
AI 모델 너깃 모델 탭.....	36
Oracle 모델 관리.....	37
Oracle 모델 너깃 서버 탭.....	37
Oracle 모델 너깃 요약 탭.....	37
Oracle 모델 너깃 설정 탭.....	37
Oracle 모델 나열.....	38
Oracle Data Miner.....	38
데이터 준비.....	38
Oracle 데이터 마이닝 예.....	39
예제 스트림: 데이터 업로드.....	39
예제 스트림: 데이터 탐색.....	40
예제 스트림: 모델 작성.....	40
예제 스트림: 모델 평가.....	40
예제 스트림: 모델 배포.....	40

## 제 5 장 IBM Data Warehouse 및 IBM Netezza Analytics를 사용한 데이터베이스 모델링. 41

IBM Data Warehouse 및 IBM Netezza Analytics가 포함된 SPSS Modeler .....	41
통합 요구사항.....	41
통합 사용.....	42
IBM Netezza Analytics 또는 IBM Data Warehouse 구성.....	42
IBM Netezza Analytics에 대한 ODBC 소스 작성 .....	42
SPSS Modeler에서 통합 사용 .....	44
SQL 생성 및 최적화 사용.....	44
IBM Netezza Analytics 및 IBM Data Warehouse를 사용한 모델 작성.....	44
필드 옵션.....	45
서버 옵션.....	46
모델 옵션.....	46
모델 관리.....	46
데이터베이스 모델 나열.....	46
IBM Data WH 회귀분석 트리.....	47
IBM Data WH 회귀분석 트리 작성 옵션 - 트리 성장.....	47
IBM Data WH 트리 작성 옵션 - 트리 가지치기.....	47
Netezza 분열 군집.....	48
Netezza 분열 군집 필드 옵션.....	48
Netezza 분열 군집 작성 옵션.....	49
IBM Data WH 일반화 선형.....	49
IBM Data WH 일반화 선형 모델 필드 옵션.....	49
IBM Data WH 일반화 선형 모델 옵션 - 일반.....	50
IBM Data WH 일반화 선형 모델 옵션 - 상호작용.....	50
IBM Data WH 일반화 선형 모델 옵션 - 스코어링 옵션.....	51
IBM Data WH 의사결정 트리.....	51
인스턴스 가중치 및 클래스 가중치.....	51
Netezza 의사결정 트리 필드 옵션.....	52
IBM Data WH 의사결정 트리 작성 옵션.....	53
IBM Data WH 선형 회귀.....	54
IBM Data WH 선형 회귀 작성 옵션.....	54
IBM Data WH KNN.....	54
IBM Data WH KNN 모델 옵션 - 일반.....	55
IBM Data WH KNN 모델 옵션 - 스코어링 옵션.....	55
IBM Data WH K-평균.....	56
IBM Data WH K-평균 필드 옵션.....	56
IBM Data WH K-평균 작성 옵션 탭.....	56
IBM Data WH Naive Bayes.....	57
Netezza Bayes Net.....	57
Netezza Bayes Net 필드 옵션.....	57
Netezza Bayes Net 작성 옵션.....	57
Netezza 시계열.....	57
Netezza 시계열에서 값의 보간법.....	58
Netezza 시계열 필드 옵션.....	60
Netezza 시계열 작성 옵션.....	60
Netezza 시계열 모델 옵션.....	62
IBM Data WH TwoStep.....	62
IBM Data WH TwoStep 필드 옵션.....	62
IBM Data WH TwoStep 작성 옵션.....	62
IBM Data WH PCA.....	63
IBM Data WH PCA 필드 옵션.....	63
IBM Data WH PCA 작성 옵션.....	63
IBM Data WH 및 Netezza 모델 관리.....	64
IBM Data Warehouse 및 IBM Netezza Analytics 모델 스코어링.....	64
IBM Data WH 및 Netezza 모델 너깃 서버 탭.....	64
IBM Data WH 의사결정 트리 모델 너깃.....	64

IBM Data WH K-평균 모델 너깃.....	65
Netezza Bayes Net 모델 너깃.....	66
IBM Data WH Naive Bayes 모델 너깃.....	67
IBM Data WH KNN 모델 너깃.....	67
Netezza 분열 군집 모델 너깃.....	68
IBM Data WH PCA 모델 너깃.....	69
Netezza 회귀분석 트리 모델 너깃.....	69
IBM Data WH 선형 회귀 모델 너깃.....	70
Netezza 시계열 모델 너깃.....	70
IBM Data WH 일반화 선형 모델 너깃.....	71
IBM Data WH TwoStep 모델 너깃.....	72
<b>제 6 장 z/OS용 IBM Db2를 사용한 데이터베이스 모델링.....</b>	<b>73</b>
IBM SPSS Modeler 및 z/OS용 IBM Db2.....	73
z/OS용 IBM Db2와의 통합을 위한 요구사항.....	73
z/OS용 IBM Db2 Analytics Accelerator와의 통합 사용.....	73
z/OS용 IBM Db2 및 z/OS용 IBM Analytics Accelerator 구성.....	74
z/OS용 IBM Db2 및 IBM Db2 Analytics Accelerator에 대한 ODBC 소스 작성.....	74
IBM SPSS Modeler에서 z/OS용 IBM Db2의 통합 사용.....	74
SQL 생성 및 최적화 사용.....	75
IBM SPSS Modeler에서 IBM Db2 클라이언트를 사용하여 DSN 구성.....	75
z/OS용 IBM Db2를 사용한 모델 작성.....	75
z/OS용 IBM Db2 모델 - 필드 옵션.....	76
z/OS용 IBM Db2 모델 - 서버 옵션.....	77
z/OS용 IBM Db2 모델 - 모델 옵션.....	77
z/OS용 IBM Db2 모델 - K-평균.....	77
z/OS용 IBM Db2 모델 - K-평균 필드 옵션.....	77
z/OS용 IBM Db2 모델 - K-평균 작성 옵션.....	77
z/OS용 IBM Db2 모델 - Naive Bayes.....	78
z/OS용 IBM Db2 모델 - 의사결정 트리.....	78
z/OS용 IBM Db2 모델 - 의사결정 트리 필드 옵션.....	78
z/OS용 IBM Db2 모델 - 의사결정 트리 작성 옵션.....	79
z/OS용 IBM Db2 모델 - 의사결정 트리 노드 - 클래스 가중치.....	79
z/OS용 IBM Db2 모델 - 의사결정 트리 노드 - 트리 가지치기.....	79
z/OS용 IBM Db2 모델 - 회귀분석 트리.....	80
z/OS용 IBM Db2 모델 - 회귀분석 트리 작성 옵션 - 트리 성장.....	80
z/OS용 IBM Db2 모델 - 회귀분석 트리 작성 옵션 - 트리 가지치기.....	80
z/OS용 IBM Db2 모델 - TwoStep.....	81
z/OS용 IBM Db2 모델 - TwoStep 필드 옵션.....	81
z/OS용 IBM Db2 모델 - TwoStep 작성 옵션.....	81
z/OS용 IBM Db2 모델 - TwoStep 너깃 - 모델 탭.....	82
IBM Db2 for z/OS 모델 관리.....	82
IBM Db2 for z/OS 모델 스코어링.....	82
z/OS용 IBM Db2 의사결정 트리 모델 너깃.....	82
z/OS용 IBM Db2 K-평균 모델 너깃.....	83
z/OS용 IBM Db2 Naive Bayes 모델 너깃.....	83
z/OS용 IBM Db2 회귀분석 트리 모델 너깃.....	83
z/OS용 IBM Db2 TwoStep 모델 너깃.....	84
<b>주의사항 .....</b>	<b>85</b>
상표.....	86
제품 문서의 이용 약관.....	86
<b>색인.....</b>	<b>89</b>

## 서론

---

IBM SPSS Modeler는 IBM 엔터프라이즈 중심의 데이터 마이닝 워크벤치입니다. SPSS Modeler는 상세한 데이터 이해를 통해 조직이 고객과 시민과의 관계를 향상시킬 수 있도록 도움을 줍니다. 조직은 SPSS Modeler에서 확보한 통찰력을 통해 수익 창출이 가능한 고객을 보유하고, 교차 판매 기회를 식별하고, 새 고객을 모으고, 사기 행위를 적발하고, 위험을 줄이고, 정부 서비스 지원을 향상시킬 수 있습니다.

SPSS Modeler의 시각적 인터페이스를 통해 사용자는 보다 쉽게 비즈니스에 특정한 전문 지식을 적용할 수 있으므로, 더 강력한 예측 모델을 생성하고 솔루션 출시 시점을 단축할 수 있습니다. SPSS Modeler에서는 예측, 분류, 세분화, 연관 발견 알고리즘과 같은 많은 모델링 기법을 제공합니다. 모델을 작성하면 IBM SPSS Modeler Solution Publisher에서 의사결정자 또는 데이터베이스까지 엔터프라이즈 범위로 모델을 전달할 수 있습니다.

### IBM Business Analytics 소개

IBM Business Analytics 소프트웨어는 의사 결정자가 비즈니스 성능을 개선하기 위해 신뢰하는 완벽하고 일관되며 정확한 정보를 제공합니다. 비즈니스 인텔리전스, 예측 분석, 금융 성과와 전략 관리 및 분석 애플리케이션의 종합 포트폴리오는 현재 성과와 앞으로의 결과를 예측하는 능력에 분명하고 즉각적이면서 실행 가능한 통찰력을 제공합니다. 풍부한 업계 솔루션, 입증된 사례 및 전문 서비스가 결합되어 어떠한 크기의 조직이라도 생산성을 극대화하고 자신있는 자동 결정을 내릴 수 있으며 더 나은 결과를 가져올 수 있습니다.

이 포트폴리오의 일부인 IBM SPSS Predictive Analytics 소프트웨어를 통해 조직은 미래의 사건을 예측하고 더 나은 비즈니스 결과를 얻기 위한 통찰력에 대해 적극적인 조치를 할 수 있습니다. 전 세계의 기업, 정부 및 학계 고객들은 고객을 매료시키고 유지하며 성장하게 만드는 동시에 불공정 행위를 줄이고 위험을 낮추는 IBM SPSS 기술의 경쟁 이점을 활용합니다. 일상 업무에서 IBM SPSS 소프트웨어를 활용한다면 예측형 기업으로 거듭날 수 있습니다. 즉 비즈니스 목표 달성을 위해 의사 결정의 방향을 정하고 이를 자동화하며 측정 가능한 경쟁 우위를 달성할 수 있습니다. 자세한 내용을 보거나 담당자에게 문의하려면 <http://www.ibm.com/spss> 사이트를 방문하십시오.

### 기술 지원

유지보수 고객은 기술 지원을 받을 수 있습니다. IBM 제품 사용 및 지원되는 하드웨어 환경 중 하나에 대해 설치하는 데 도움이 필요한 경우 기술 지원에 문의하십시오. 기술 지원에 문의하려면 IBM 웹 사이트(<http://www.ibm.com/support>)를 참조하십시오. 지원을 요청하려면 본인의 신상과 소속 조직(회사) 및 지원 동의서를 제시해야 합니다.



# 제 1 장 IBM SPSS Modeler 정보

IBM SPSS Modeler는 비즈니스 전문 지식을 사용하여 예측 모형을 신속하게 개발하고 이를 비즈니스 운영에 배포하여 의사결정의 정확성을 향상시켜주는 데이터 마이닝 도구 세트입니다. 산업 표준 CRISP-DM 모델을 중심으로 디자인된 IBM SPSS Modeler는 데이터에서 보다 나아진 비즈니스 결과에 이르는 전체 데이터 마이닝 프로세스를 지원합니다.

IBM SPSS Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다. 모델링 팔레트에서 사용할 수 있는 이러한 방법을 통해 데이터로부터 새로운 정보를 얻어서 예측 모형을 개발할 수 있습니다. 각각의 방법은 그것만의 장점이 있으며 특정한 문제점 유형에 가장 적합합니다.

SPSS Modeler는 독립형 제품으로 구매하거나 SPSS Modeler Server와 통합하여 클라이언트로 사용할 수 있습니다. 다음 절에 요약된 바와 같이 여러가지 추가 옵션도 사용할 수 있습니다. 자세한 정보는 <https://www.ibm.com/analytics/us/en/technology/spss/>의 내용을 참조하십시오.

## IBM SPSS Modeler 제품

IBM SPSS Modeler 제품군 및 연관 소프트웨어는 다음으로 구성됩니다.

- IBM SPSS Modeler
- IBM SPSS Modeler Server
- IBM SPSS Modeler Administration Console(IBM SPSS Deployment Manager와 함께 포함)
- IBM SPSS Modeler Batch
- IBM SPSS Modeler Solution Publisher
- IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

## IBM SPSS Modeler

SPSS Modeler는 개인용 컴퓨터에 설치하여 실행되는 기능적으로 완벽한 버전의 제품입니다. 로컬 모드에서 독립형 제품으로 SPSS Modeler를 실행하거나 대형 데이터 세트에 대한 성능 향상을 위해 분산 모드에서 IBM SPSS Modeler Server와 함께 사용할 수 있습니다.

SPSS Modeler를 사용하여 프로그래밍하지 않고 신속하게 직관적으로 정확한 예측 모델을 작성할 수 있습니다. 고유한 시각적 인터페이스를 사용하면 데이터 마이닝 프로세스를 쉽게 시각화할 수 있습니다. 제품에 포함된 고급 분석 지원을 통해 데이터에서 이전에 숨겨진 패턴과 추세를 발견할 수 있습니다. 결과를 모델링하고 결과에 영향을 주는 요인을 이해하여 비즈니스 기회를 활용하고 위험을 줄일 수 있습니다.

SPSS Modeler는 두 개의 에디션(SPSS Modeler Professional과 SPSS Modeler Premium)으로 사용할 수 있습니다. 자세한 정보는 2 페이지의 [『IBM SPSS Modeler 에디션』](#)의 내용을 참조하십시오.

## IBM SPSS Modeler Server

SPSS Modeler는 클라이언트/서버 아키텍처를 사용하여 자원 집약적 작업에 대한 요청을 강력한 서버 소프트웨어로 분배하여 대형 데이터 세트에 대한 성능을 향상시킵니다.

SPSS Modeler Server는 하나 이상의 IBM SPSS Modeler 설치와 함께 서버 호스트의 분산 분석 모드에서 계속해서 실행되는 별도로 라이센스가 부여된 제품입니다. 이런 방법으로 클라이언트 컴퓨터로 데이터를 다운로드하지 않고 서버에서 메모리 집약적 작업을 수행할 수 있기 때문에 SPSS Modeler Server는 대형 데이터 세트에 대한 우수한 성능을 제공합니다. 또한 IBM SPSS Modeler Server는 SQL 최적화 및 In-Database 모델링 기능에 대한 지원을 제공하여 성능 및 자동화의 이점도 추가로 제공합니다.

## IBM SPSS Modeler Administration Console

Modeler Administration Console은 옵션 파일을 통해서도 구성 가능한 다수의 SPSS Modeler Server 구성 옵션을 관리하기 위한 그래픽 사용자 인터페이스입니다. 콘솔은 IBM SPSS Deployment Manager에 포함되며, SPSS Modeler Server 설치를 모니터하고 구성하는 데 사용할 수 있고, 현재 SPSS Modeler Server 고객이 무료로 사용

할 수 있습니다. 이 애플리케이션은 Windows 컴퓨터에만 설치할 수 있지만 지원되는 플랫폼에 설치된 서버를 관리할 수 있습니다.

## IBM SPSS Modeler Batch

데이터 마이닝은 일반적으로 대화식 처리인 반면, 그래픽 사용자 인터페이스가 없어도 명령행에서 SPSS Modeler를 실행할 수 있습니다. 예를 들어, 사용자 개입 없이 수행할 장기 실행 또는 반복 작업이 있습니다. SPSS Modeler Batch는 정규 사용자 인터페이스에 대한 액세스 없이 SPSS Modeler의 전체 분석 기능에 대한 지원을 제공하는 특수 버전의 제품입니다. SPSS Modeler Batch를 사용하려면 SPSS Modeler Server가 필요합니다.

## IBM SPSS Modeler Solution Publisher

SPSS Modeler Solution Publisher는 외부 런타임 엔진을 통해 실행하거나 외부 애플리케이션에 포함될 수 있는 SPSS Modeler 스트림의 패키지 버전을 작성할 수 있게 하는 도구입니다. 이런 방법으로 SPSS Modeler가 설치되지 않는 환경에 사용할 수 있도록 전체 SPSS Modeler 스트림을 출판하고 배포할 수 있습니다. SPSS Modeler Solution Publisher는 별도의 라이센스가 필요한 IBM SPSS Collaboration and Deployment Services - Scoring 서비스의 일부로 분배됩니다. 이 라이센스가 있으면 출판된 스트림을 실행할 수 있게 하는 SPSS Modeler Solution Publisher Runtime을 수신합니다.

SPSS Modeler Solution Publisher에 대한 자세한 정보는 IBM SPSS Collaboration and Deployment Services 문서를 참조하십시오. IBM SPSS Collaboration and Deployment Services IBM Documentation에는 "IBM SPSS Modeler Solution Publisher" 및 "IBM SPSS Analytics Toolkit" 섹션이 포함되어 있습니다.

## IBM SPSS Collaboration and Deployment Services용 IBM SPSS Modeler Server 어댑터

SPSS Modeler와 SPSS Modeler Server가 IBM SPSS Collaboration and Deployment Services 리포지토리와 상호작용할 수 있게 하는 IBM SPSS Collaboration and Deployment Services용 어댑터를 상당수 사용할 수 있습니다. 이런 방법으로 리포지토리에 배포된 SPSS Modeler 스트림을 여러 사용자가 공유하거나 씬 클라이언트 애플리케이션 IBM SPSS Modeler Advantage에서 액세스할 수 있습니다. 리포지토리를 호스팅하는 시스템에 어댑터를 설치하십시오.

## IBM SPSS Modeler 에디션

---

SPSS Modeler는 다음 에디션으로 사용할 수 있습니다.

### SPSS Modeler Professional

SPSS Modeler Professional은 CRM 시스템, 인구 통계, 구매 동작, 판매 데이터에서 추적된 동작 및 상호작용과 같은 대부분의 구조화된 데이터 유형에 대한 작업을 하는 데 필요한 모든 도구를 제공합니다.

### SPSS Modeler Premium

SPSS Modeler Premium은 특수 데이터 및 비정형 텍스트 데이터에 대한 작업을 할 수 있도록 SPSS Modeler Professional을 확장하는 별도로 라이센스가 부여된 제품입니다. SPSS Modeler Premium에는 IBM SPSS Modeler Text Analytics가 포함됩니다.

**IBM SPSS Modeler Text Analytics**는 고급 언어 기술 및 자연어 처리(NLP)를 사용하여 다양한 비정형 텍스트 데이터를 빠르게 처리하고, 주요 개념을 추출 및 구성하며, 이러한 개념을 범주로 분류합니다. 추출된 개념과 범주는 인구 통계와 같은 기준의 구조화된 데이터와 결합할 수 있고 보다 나은 집중적인 의사결정을 내리기 위해 전체 IBM SPSS Modeler 데이터 마이닝 세트를 사용하여 모델링에 적용할 수 있습니다.

### IBM SPSS Modeler Subscription

IBM SPSS Modeler Subscription에서는 일반적인 IBM SPSS Modeler 클라이언트와 동일한 모든 예측 분석 공정 능력을 제공합니다. Subscription 에디션을 사용하면 제품 업데이트를 정기적으로 다운로드할 수 있습니다.

# 문서

---

문서는 SPSS Modeler의 도움말 메뉴에서 사용할 수 있습니다. 이는 제품 외부에서 항상 사용 가능한 온라인 IBM Documentation을 엽니다.

설치 지시사항을 포함하여 각 제품에 대한 전체 문서 또한 제품 다운로드의 일부로 별도의 압축 폴더에 PDF 형식으로 제공됩니다. 또는 웹(<https://www.ibm.com/support/pages/spss-modeler-183-documentation>)에서 최신 PDF 문서를 다운로드할 수 있습니다.

## SPSS Modeler Professional 문서

SPSS Modeler Professional 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **IBM SPSS Modeler 사용자 안내서.** SPSS Modeler 사용에 관한 일반적인 소개이며, 데이터 스트림 작성, 결측 값 처리, CLEM 표현식 작성 프로젝트 및 보고서에 대한 작업, IBM SPSS Collaboration and Deployment Services 또는 IBM SPSS Modeler Advantage에 배포하기 위한 스트림 패키지 방법이 포함됩니다.
- **IBM SPSS Modeler 소스, 프로세스 및 출력 노드.** 여러 형식의 데이터를 읽고 처리하며, 출력하는 데 사용하는 모든 노드에 대한 설명입니다. 실질적으로 이는 모델링 노드 이외의 모든 노드를 의미합니다.
- **IBM SPSS Modeler 모델링 노드.** 데이터 마이닝 모델을 작성하는 데 사용하는 모든 노드에 대한 설명입니다. IBM SPSS Modeler는 기계 학습, 인공지능 및 통계로부터 취한 다양한 모델링 방법을 제공합니다.
- **IBM SPSS Modeler 애플리케이션 안내서.** 이 안내서의 예제는 특정 모델링 방법과 기법을 중점적으로 간략히 소개합니다. 이 안내서의 온라인 버전은 도움말 메뉴에서도 사용할 수 있습니다. 자세한 정보는 [3 페이지의 『애플리케이션 예제』](#) 주제를 참조하십시오.
- **IBM SPSS Modeler Python 스크립팅 및 자동화.** 노드와 스트림을 조작하는 데 사용할 수 있는 특성을 포함하여 Python 스크립팅을 통한 시스템 자동화에 대한 정보입니다.
- **IBM SPSS Modeler 배포 안내서.** IBM SPSS Deployment Manager에서 작업 처리 단계로 IBM SPSS Modeler 스트림 실행에 대한 정보입니다.
- **IBM SPSS Modeler CLEF 개발자 안내서.** CLEF는 데이터 처리 루틴 또는 모델링 알고리즘과 같은 써드파티 프로그램을 IBM SPSS Modeler의 노드로 통합하는 기능을 제공합니다.
- **IBM SPSS Modeler In-Database 마이닝 안내서.** 데이터베이스의 능력을 사용하여 성능을 향상시키고 써드 파티 알고리즘을 통해 분석 기능 범위를 확장하는 방법에 대한 정보입니다.
- **IBM SPSS Modeler Server 관리 및 성능 안내서.** IBM SPSS Modeler Server 구성 및 관리 방법에 대한 정보입니다.
- **IBM SPSS Deployment Manager 사용자 안내서.** IBM SPSS Modeler Server 모니터링 및 구성용 Deployment Manager 애플리케이션에 포함된 관리 콘솔 사용자 인터페이스를 사용하는 데 대한 정보
- **IBM SPSS Modeler CRISP-DM 안내서.** SPSS Modeler에서 데이터 마이닝에 CRISP-DM 방법론을 사용하기 위한 단계별 안내서입니다.
- **IBM SPSS Modeler Batch 사용자 안내서.** 일괄처리 모드 실행 및 명령행 인수 세부사항을 포함하여 일괄처리 모드에서 IBM SPSS Modeler 사용을 위한 전체 안내서입니다. 이 안내서는 PDF 형식으로만 사용할 수 있습니다.

## SPSS Modeler Premium 문서

SPSS Modeler Premium 문서 스위트(설치 지시사항은 제외)는 다음과 같습니다.

- **SPSS Modeler Text Analytics 사용자 안내서.** SPSS Modeler에서 텍스트 분석 사용에 대한 정보로, 텍스트 마이닝 노드, 대화식 워크벤치, 템플리트 및 기타 자원에 대해 설명합니다.

## 애플리케이션 예제

---

SPSS Modeler의 데이터 마이닝 도구가 광범위한 비즈니스 및 조직의 문제점을 해결하는 데 도움을 주는 가운데, 애플리케이션 예제는 특정 모델링 방법 및 기술에 대해 대상화된 간략한 소개를 제공합니다. 여기서 사용된 데이터 세트는 일부 데이터 마이너에서 관리하는 거대한 데이터 스토어보다 훨씬 작지만, 관련된 개념과 방법은 실제 애플리케이션으로 확장 가능합니다.

예제에 액세스하려면 SPSS Modeler의 도움말 메뉴에서 **애플리케이션 예제**를 클릭하십시오.

데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래에 있는 Demos 폴더에 설치됩니다. 추가 정보는 4 페 이지의 [『Demos 폴더』](#)의 내용을 참조하십시오.

**데이터베이스 모델링 예제.** IBM SPSS Modeler In-Database 마이닝 안내서의 예제를 참조하십시오.

**스크립팅 예제.** IBM SPSS Modeler 스크립팅 및 자동화 안내서의 예제를 참조하십시오.

## Demos 폴더

---

애플리케이션 예제에서 사용하는 데이터 파일 및 샘플 스트림은 제품 설치 디렉토리 아래의 Demos 폴더(예: C:\Program Files\IBM\SPSS\Modeler\<version>\Demos)에 설치됩니다. Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서, 또는 파일 > 스트림 열기 대화 상자의 최근 디렉토리 목록에서 Demos를 클릭해서도 이 폴더에 액세스할 수 있습니다.

## 라이센스 추적

---

SPSS Modeler를 사용할 때 라이센스 사용이 정기적으로 추적되고 로그됩니다. 로그되는 라이센스 메트릭은 AUTHORIZED\_USER 및 CONCURRENT\_USER이며 로그되는 메트릭의 유형은 SPSS Modeler에 대해 가진 라이센스의 유형에 의해 결정됩니다.

생성되는 로그 파일은 사용자가 라이센스 사용 보고서를 생성할 수 있는 IBM 라이센스 메트릭 도구에 의해 처리될 수 있습니다.

라이센스 로그 파일은 SPSS Modeler 클라이언트 로그 파일이 기록되는 디렉토리와 동일한 디렉토리(기본적으로 %ALLUSERSPROFILE%\IBM\SPSS\Modeler\<version>/log)에 작성됩니다.

## 제 2 장 In-Database 마이닝

### 데이터베이스 모델링 개요

IBM SPSS Modeler Server는 IBM Netezza, Oracle Data Miner, Microsoft Analysis Services를 포함한 데이터베이스 벤더로부터 사용할 수 있는 데이터 마이닝 및 모델링 도구와의 통합을 지원합니다. 모두 IBM SPSS Modeler 애플리케이션 내에서 시작하여 데이터베이스 내에 모델을 작성하고, 스코어링하고, 저장할 수 있습니다. 이를 통해 이 벤더가 제공하는 데이터베이스 원시 알고리즘을 활용하면서 데이터베이스의 성능과 IBM SPSS Modeler의 분석 기능 및 편리한 사용을 결합할 수 있습니다. 모델은 데이터베이스 내부에서 작성되므로 필요한 경우 일반적인 방식으로 IBM SPSS Modeler 인터페이스를 통해 찾아서 스코어링한 후 IBM SPSS Modeler Solution Publisher를 사용하여 배포할 수 있습니다. 지원되는 알고리즘은 IBM SPSS Modeler의 데이터베이스 모델링 팔레트에 있습니다.

IBM SPSS Modeler를 사용하여 데이터베이스 원시 알고리즘에 액세스하면 다음과 같은 여러 가지 장점이 있습니다.

- In-Database 알고리즘은 종종 데이터베이스 서버와 밀접하게 통합되어 향상된 성능을 제공할 수 있습니다.
- "데이터베이스에서" 작성되고 스코어링된 모델은 데이터베이스에 액세스할 수 있는 애플리케이션에 쉽게 배치하고 이 애플리케이션과 공유할 수 있습니다.

**SQL 생성.** In-Database 모델링은 "SQL 푸시백"이라고도 알려져 있는 SQL 생성과 구별됩니다. 이 기능을 사용하면 성능 향상을 위해 데이터베이스에 "푸시백"(즉, 데이터베이스에서 실행)할 수 있는 원시 IBM SPSS Modeler 조작에 대한 SQL문을 생성할 수 있습니다. 예를 들어, 병합, 통합 및 선택 노드는 모두 이 방식으로 데이터베이스에 푸시백할 수 있는 SQL 코드를 생성합니다. 데이터베이스 모델링과 함께 SQL 생성을 사용하면 데이터베이스의 시작부터 끝까지 실행될 수 있는 스트림이 생성되어 IBM SPSS Modeler에서 실행되는 스트림보다 상당히 성능이 향상됩니다.

**참고:** 데이터베이스 모델링 및 SQL 최적화를 위해서는 IBM SPSS Modeler 컴퓨터에서 IBM SPSS Modeler Server 연결이 사용으로 설정되어 있어야 합니다. 이 설정이 사용 가능하면 데이터베이스 알고리즘에 액세스하고, SQL을 IBM SPSS Modeler에서 직접 푸시백하고, IBM SPSS Modeler Server에 액세스할 수 있습니다. 현재 라이센스 상태를 검증하려면 IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

#### 도움말 > 정보 > 추가 세부사항

연결성이 사용 가능으로 설정되면 라이센스 상태 탭에 **서버 사용 가능** 옵션이 표시됩니다.

지원되는 알고리즘에 대한 정보는 특정 벤더에 대한 후속 절을 참조하십시오.

### 필요사항

데이터베이스 모델링을 수행하려면 다음의 설정이 필요합니다.

- 필수 분석 구성요소(Microsoft Analysis Services 또는 Oracle Data Miner)가 설치된 적절한 데이터베이스에 대한 ODBC 연결
- IBM SPSS Modeler에서는 헬퍼 애플리케이션 대화 상자(도구 > 헬퍼 애플리케이션)에서 데이터베이스 모델링을 사용으로 설정해야 합니다.
- IBM SPSS Modeler 및 IBM SPSS Modeler Server(사용된 경우)의 사용자 옵션 대화 상자에서 **SQL 생성** 및 **SQL 최적화** 설정을 사용으로 설정해야 합니다. SQL 최적화는 데이터베이스 모델링이 작동하기 위해 엄격하게 요구되지 않지만 성능상 이유로 강력하게 권장됩니다.

**참고:** 데이터베이스 모델링 및 SQL 최적화를 위해서는 IBM SPSS Modeler 컴퓨터에서 IBM SPSS Modeler Server 연결이 사용으로 설정되어 있어야 합니다. 이 설정이 사용 가능하면 데이터베이스 알고리즘에 액세스하고, SQL을 IBM SPSS Modeler에서 직접 푸시백하고, IBM SPSS Modeler Server에 액세스할 수 있습니다. 현재 라이센스 상태를 검증하려면 IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

#### 도움말 > 정보 > 추가 세부사항

연결성이 사용 가능으로 설정되면 라이센스 상태 탭에 **서버 사용 가능** 옵션이 표시됩니다.

자세한 정보는 특정 벤더에 대한 후속 절을 참조하십시오.

## 모델 작성

데이터베이스 알고리즘을 사용하여 모델을 작성하고 스코어링하는 프로세스는 IBM SPSS Modeler에서 다른 유형의 데이터 마이닝과 비슷합니다. 노드에 대해 작업하고 "너깃"을 모델링하는 일반적인 프로세스는 IBM SPSS Modeler에서 작업할 때의 다른 스트림과 비슷합니다. 유일한 차이점은 실제 처리 및 모델 작성이 데이터베이스에 푸시백되는 것입니다.

데이터베이스 모델링 스트림은 개념적으로 IBM SPSS Modeler의 다른 데이터 스트림과 동일하지만 이 스트림은 Microsoft 의사결정 트리 노드를 사용한 모델 작성 등을 포함한 모든 조작을 데이터베이스에서 수행합니다. 스트림을 실행하면 IBM SPSS Modeler가 결과 모델을 작성하고 저장하도록 데이터베이스에 지시하며 세부사항이 IBM SPSS Modeler에 다운로드됩니다. In-Database 실행은 스트림에서 보라색 음영 처리된 노드를 사용하여 표시됩니다.

## 데이터 준비

데이터베이스 원시 알고리즘이 사용되는지 여부에 관계없이 성능을 향상시키기 위해 가능할 때마다 데이터 준비를 데이터베이스에 푸시백해야 합니다.

- 원래 데이터가 데이터베이스에 저장되는 경우 목표는 모든 필수 업스트림 조작을 SQL로 변환될 수 있게 하여 해당 데이터를 데이터베이스에서 유지하는 것입니다. 그러면 데이터가 IBM SPSS Modeler에 다운로드되는 것을 방지하여 성능 향상을 무효화하는 병목 현상을 피하고 전체 스트림을 데이터베이스에서 실행할 수 있습니다.
- 원래 데이터가 데이터베이스에 저장되지 않는 경우에는 데이터베이스 모델링을 계속 사용할 수 있습니다. 이 경우 데이터 준비는 IBM SPSS Modeler에서 수행되며 준비된 데이터 세트는 모델 작성을 위해 자동으로 데이터베이스에 업로드됩니다.

## 모델 스코어링

In-Database 마이닝을 사용하여 IBM SPSS Modeler에서 생성된 모델은 일반적인 IBM SPSS Modeler 모델과 다릅니다. 해당 모델은 생성된 모델 "너깃"으로 모델 관리자에 표시되지만 실제로는 원격 데이터 마이닝 또는 데이터베이스 서버에 보유되는 원격 모델입니다. IBM SPSS Modeler에 표시되는 것은 단순히 이 원격 모델에 대한 참조입니다. 즉, 표시되는 IBM SPSS Modeler 모델은 데이터베이스 서버 호스트 이름, 데이터베이스 이름, 모델 이름 등의 정보가 포함된 "비어 있는" 모델입니다. 이는 데이터베이스 원시 알고리즘을 사용하여 작성되는 모델을 찾아보고 스코어링할 때 이해할 중요한 차이입니다.

모델을 작성한 후에는 IBM SPSS Modeler의 생성된 다른 모델과 마찬가지로 스코어링을 위해 스트림에 해당 모델을 추가할 수 있습니다. 업스트림 조작은 제외하고 모든 스코어링이 데이터베이스 내에서 수행됩니다. (가능한 경우 성능 향상을 위해 업스트림 조작은 여전히 데이터베이스로 푸시백할 수 있지만 이는 스코어링 수행을 위한 요구사항은 아닙니다.) 또한 데이터베이스 벤더가 제공하는 표준 브라우저를 사용하여 대부분의 경우 생성된 모델을 찾아볼 수 있습니다.

찾아보기와 스코어링 모두에 대해 Oracle Data Miner 또는 Microsoft Analysis Services를 실행 중인 서버에 대한 활성 연결이 필요합니다.

## 결과 보기 및 설정 지정

결과를 보고 스코어링에 대한 설정을 지정하려면 스트림 캔버스에서 모델을 두 번 클릭하십시오. 또는 모델을 마우스 오른쪽 단추로 클릭한 후 찾아보기 또는 편집을 선택할 수 있습니다. 구체적인 설정은 모델 유형에 따라 달립니다.

## 데이터베이스 모델 내보내기 및 저장

데이터베이스 모델 및 요약은 파일 메뉴의 옵션을 사용하여 IBM SPSS Modeler에서 작성된 기타 모델과 동일한 방식으로 모델 브라우저에서 내보낼 수 있습니다.

- 모델 브라우저의 파일 메뉴에 있는 다음 옵션 중에서 선택하십시오.
  - 텍스트 내보내기 - 모델 요약을 텍스트 파일로 내보냄

- **HTML 내보내기** - 모델 요약을 HTML 파일로 내보냄
- **PMMI 내보내기**(IBM Db2 IM 모델의 경우에만 지원됨) - 모델을 예측 모델 마크업 언어(PMMI)로 내보내어 다른 PMML 호환 소프트웨어에서 사용할 수 있게 함

**참고:** 파일 메뉴에서 노드 저장을 선택하여 생성된 모델을 저장할 수도 있습니다.

## 모델 일관성

생성되는 각각의 데이터베이스 모델에 대해 IBM SPSS Modeler는 데이터베이스에 저장되는 동일한 이름의 모델에 대한 참조와 함께 모델 구조에 대한 설명을 저장합니다. 생성된 모델의 서버 탭에는 데이터베이스에서 실제 모델과 일치하는 해당 모델에 대해 생성된 고유 키가 표시됩니다.

IBM SPSS Modeler는 이 무작위로 생성된 키를 사용하여 모델이 일관성을 계속 유지하는지 확인합니다. 이 키는 작성될 때 모델의 설명에 저장됩니다. 배포 스트림을 실행하기 전에 키가 일치하는지 확인하는 것이 좋습니다.

1. 해당 설명을 IBM SPSS Modeler가 저장한 무작위 키와 비교하여 데이터베이스에 저장된 모델의 일관성을 확인하려면 확인 단추를 클릭하십시오. 데이터베이스 모델을 찾을 수 없거나 키가 일치하지 않으면 오류가 보고됩니다.

## 생성된 SQL 보기 및 내보내기

생성된 SQL 코드는 실행 전에 미리 볼 수 있어서 디버깅을 위해 유용할 수 있습니다.



# 제 3 장 Microsoft Analysis Services를 사용한 데이터베이스 모델링

## IBM SPSS Modeler 및 Microsoft Analysis Services

IBM SPSS Modeler는 Microsoft SQL Server Analysis Services와의 통합을 지원합니다. 이 기능은 IBM SPSS Modeler에서 모델링 노드로 구현되며 데이터베이스 모델링 팔레트에서 사용할 수 있습니다. 해당 팔레트가 표시되지 않으면 헬퍼 애플리케이션 대화 상자에서 Microsoft 탭에 있는 MS Analysis Services 통합을 사용으로 설정하여 이 기능을 활성화할 수 있습니다. 자세한 정보는 11 페이지의 『Analysis Services와의 통합 사용』의 내용을 참조하십시오.

IBM SPSS Modeler는 다음과 같은 Analysis Services 알고리즘의 통합을 지원합니다.

- 의사결정 트리
- 군집화
- 연관 규칙
- Naive Bayes
- 선형 회귀
- 신경망
- 로지스틱 회귀분석
- 시계열
- 시퀀스 군집화

다음 다이어그램에서는 클라이언트로부터 IBM SPSS Modeler Server에 의해 In-Database 마이닝이 관리되는 서버로의 데이터 플로우를 보여줍니다. 모델 작성은 Analysis Services를 사용하여 수행됩니다. 결과 모델은 Analysis Services에 의해 저장됩니다. 이 모델에 대한 참조는 IBM SPSS Modeler 스트림 내에서 유지됩니다. 그런 다음 해당 모델은 스코어링을 위해 Analysis Services로부터 Microsoft SQL Server 또는 IBM SPSS Modeler로 다운로드됩니다.

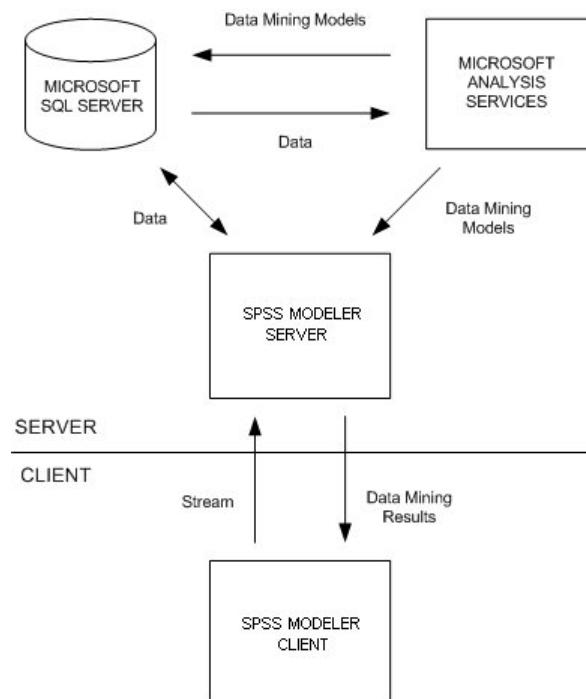


그림 1. 모델 작성 중 IBM SPSS Modeler, Microsoft SQL Server 및 Microsoft Analysis Services 간 데이터 플로우

참고: IBM SPSS Modeler Server는 필수는 아니지만 사용할 수는 있습니다. IBM SPSS Modeler 클라이언트는 In-Database 마이닝 계산을 자체적으로 처리할 수 있습니다.

## Microsoft Analysis Services와의 통합을 위한 요구사항

IBM SPSS Modeler와 함께 Analysis Services 알고리즘을 사용하여 In-Database 모델링을 수행하려면 다음과 같은 전제조건이 있습니다. 데이터베이스 관리자에게 문의하여 이 조건이 충족되는지 확인해야 할 수 있습니다.

- Windows의 IBM SPSS Modeler Server 설치(분산 모드)에 대해 실행 중인 IBM SPSS Modeler. UNIX 플랫폼은 Analysis Services와의 이 통합에서 지원되지 않습니다.

**중요사항:** IBM SPSS Modeler 사용자는 추가 IBM SPSS Modeler Server 요구사항에 나열된 URL에서 Microsoft로부터 얻을 수 있는 SQL 원시 클라이언트 드라이버를 사용하여 ODBC 연결을 구성해야 합니다. IBM SPSS Data Access Pack과 함께 제공되고 일반적으로 IBM SPSS Modeler의 기타 사용에 대해 권장되는 드라이버는 이 용도에는 권장되지 않습니다. IBM SPSS Modeler는 SQL Server 인증을 지원하지 않으므로 통합된 Windows 인증 사용이 사용으로 설정된 SQL Server를 사용하기 위해 이 드라이버를 구성해야 합니다.

ODBC 데이터 소스에 대한 작성 및 설정에 관한 문의사항이 있으면 데이터베이스 관리자에게 문의하십시오.

- SQL Server는 설치되어 있어야 하지만, 반드시 IBM SPSS Modeler와 동일한 호스트에 설치할 필요는 없습니다. IBM SPSS Modeler 사용자는 데이터를 읽고 쓰고 테이블 및 보기를 삭제하고 작성하는 데 필요한 충분한 권한을 가지고 있어야 합니다.

**참고:** SQL Server Enterprise Edition이 권장됩니다. Enterprise Edition은 알고리즘 결과를 조정하는 고급 모수를 제공하여 추가적인 유연성을 제공합니다. Standard Edition 버전은 동일한 모수를 제공하지만 사용자에게 고급 모수 중 일부의 편집을 허용하지 않습니다.

- Microsoft SQL Server Analysis Services가 SQL Server와 동일한 호스트에 설치되어 있어야 합니다.

### 추가 IBM SPSS Modeler Server 요구 사항

Analysis Services 알고리즘을 IBM SPSS Modeler Server와 함께 사용하려면 다음과 같은 구성요소가 IBM SPSS Modeler Server 호스트 시스템에 설치되어 있어야 합니다.

**참고:** SQL Server가 IBM SPSS Modeler Server와 동일한 호스트에 설치되어 있는 경우에는 이들 구성요소가 이미 설치되어 있습니다.

- Microsoft SQL Server Analysis Services 10.0 OLE DB Provider(사용자의 운영 체제에 맞는 올바른 변형을 선택해야 함)
- Microsoft SQL Server Native Client(사용자의 운영 체제에 맞는 올바른 변형을 선택해야 함)
- Microsoft SQL Server 2008 또는 2012를 사용하고 있는 경우에는 Microsoft Core XML Services(MSXML) 6.0도 필요할 수 있습니다.

이 구성요소를 다운로드하려면 [www.microsoft.com/downloads](http://www.microsoft.com/downloads)로 이동하여 **.NET Framework** 또는 **SQL Server Feature Pack**(다른 모든 구성요소의 경우)을 검색한 후 사용자의 SQL Server 버전에 맞는 최신 팩을 선택하십시오.

이를 위해서는 다른 패키지를 먼저 설치해야 할 수 있습니다(해당 패키지도 Microsoft 다운로드 웹 사이트에서 얻을 수 있음).

### 추가 IBM SPSS Modeler 요구 사항

Analysis Services 알고리즘을 IBM SPSS Modeler와 함께 사용하려면 위와 동일한 구성요소를 설치해야 하며 클라이언트에서 다음을 추가해야 합니다.

- Microsoft SQL Server Datamining Viewer Controls(사용자의 운영 체제에 맞는 올바른 변형을 선택해야 함) - 이 구성요소에는 다음 구성요소도 필요합니다.
- Microsoft ADOMD.NET

이 구성요소를 다운로드하려면 [www.microsoft.com/downloads](http://www.microsoft.com/downloads)로 이동하여 **SQL Server Feature Pack**을 검색한 후 사용자의 SQL Server 버전에 맞는 최신 팩을 선택하십시오.

**참고:** 데이터베이스 모델링 및 SQL 최적화를 위해서는 IBM SPSS Modeler 컴퓨터에서 IBM SPSS Modeler Server 연결이 사용으로 설정되어 있어야 합니다. 이 설정이 사용 가능하면 데이터베이스 알고리즘에 액세스하

고, SQL을 IBM SPSS Modeler에서 직접 푸시백하고, IBM SPSS Modeler Server에 액세스할 수 있습니다. 현재 라이센스 상태를 검증하려면 IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

#### 도움말 > 정보 > 추가 세부사항

연결성이 사용 가능으로 설정되면 라이센스 상태 탭에 **서버 사용 가능** 옵션이 표시됩니다.

## Analysis Services와의 통합 사용

Analysis Services와 IBM SPSS Modeler의 통합을 사용하려면 SQL Server 및 Analysis Services를 구성하고 ODBC 소스를 작성하고 IBM SPSS Modeler 헬퍼 애플리케이션 대화 상자에서 통합을 사용으로 설정하고 SQL 생성 및 최적화를 사용으로 설정해야 합니다.

참고: Microsoft SQL Server 및 Microsoft Analysis Services를 사용할 수 있어야 합니다. 자세한 정보는 [10 페이지의 『Microsoft Analysis Services와의 통합을 위한 요구사항』](#)의 내용을 참조하십시오.

#### SQL Server 구성

데이터베이스 내에서 스코어링이 수행될 수 있도록 SQL Server를 구성하십시오.

1. SQL Server 호스트 시스템에서 다음 레지스트리 키를 작성하십시오.

```
HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\MSSQLServer\Providers\MSOLAP
```

2. 다음 DWORD 값을 이 키에 추가하십시오.

```
AllowInProcess 1
```

3. 이 변경사항을 작성한 후 SQL Server를 다시 시작하십시오.

#### Analysis Services 구성

IBM SPSS Modeler가 Analysis Services와 통신하려면 먼저 분석 서버 특성 대화 상자에서 두 가지 설정을 수동으로 구성해야 합니다.

1. MS SQL Server Management Studio를 통해 분석 서버에 로그인하십시오.
2. 서버 이름을 마우스 오른쪽 단추로 클릭한 후 **특성**을 선택하여 특성 대화 상자에 액세스하십시오.
3. **고급(모든) 특성 표시** 선택란을 선택하십시오.
4. 다음과 같은 특성을 변경하십시오.
  - DataMining\AllowAdHocOpenRowsetQueries의 값을 True로 변경하십시오(기본값은 False임).
  - DataMining\AllowProvidersInOpenRowset의 값을 [all]로 변경하십시오(기본값이 없음).

#### SQL Server에 대한 ODBC DSN 작성

데이터베이스를 읽거나 데이터베이스에 쓰려면 필요에 따라 읽기 또는 쓰기 권한을 가지고 관련 데이터베이스에 대해 ODBC 데이터 소스가 설치 및 구성되어 있어야 합니다. Microsoft SQL 원시 클라이언트 ODBC 드라이버는 필수이며 SQL Server와 함께 자동으로 설치됩니다. IBM SPSS Data Access Pack과 함께 제공되고 일반적으로 IBM SPSS Modeler의 기타 사용에 대해 권장되는 드라이버는 이 용도에는 권장되지 않습니다. IBM SPSS Modeler와 SQL Server가 서로 다른 호스트에 상주하는 경우에는 Microsoft SQL 원시 클라이언트 ODBC 드라이버를 다운로드할 수 있습니다. 자세한 정보는 [10 페이지의 『Microsoft Analysis Services와의 통합을 위한 요구사항』](#)의 내용을 참조하십시오.

ODBC 데이터 소스에 대한 작성 및 설정에 관한 문의사항이 있으면 데이터베이스 관리자에게 문의하십시오.

1. Microsoft SQL 원시 클라이언트 ODBC 드라이버를 사용하여 데이터 마이닝 프로세스에서 사용되는 SQL Server 데이터베이스를 가리키는 ODBC DSN을 작성하십시오. 나머지 기본 드라이버 설정을 사용해야 합니다.
2. 이 DSN의 경우 **통합된 Windows 인증 사용**이 선택되어 있는지 확인하십시오.
  - IBM SPSS Modeler와 IBM SPSS Modeler Server가 서로 다른 호스트에서 실행 중인 경우에는 각각의 호스트에서 동일한 ODBC DSN을 작성하십시오. 각 호스트에서 동일한 DSN 이름이 사용되는지 확인하십시오.

#### IBM SPSS Modeler에서 Analysis Services 통합 사용

IBM SPSS Modeler가 Analysis Services를 사용할 수 있게 하려면 먼저 헬퍼 애플리케이션 대화 상자에서 서버 사양을 제공해야 합니다.

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

도구 > 옵션 > 헬퍼 애플리케이션

2. Microsoft 탭을 클릭하십시오.

- **Microsoft Analysis Services 통합을 사용으로 설정하십시오.** IBM SPSS Modeler 창의 맨 아래에서 데이터베이스 모델링 팔레트를 사용으로 설정하고(아직 표시되지 않은 경우) Analysis Services 알고리즘에 대한 노드를 추가합니다.
- **분석 서버 호스트.** Analysis Services가 실행 중인 시스템의 이름을 지정하십시오.
- **분석 서버 데이터베이스.** 생략 기호(...) 단추를 클릭하여 사용 가능한 데이터베이스 중에서 선택할 수 있는 하위 대화 상자를 열어 원하는 데이터베이스를 선택하십시오. 목록은 지정된 분석 서버에 사용 가능한 데이터베이스로 채워져 있습니다. Microsoft Analysis Services는 데이터 마이닝 모델을 이름 지정된 데이터베이스에 저장하므로 IBM SPSS Modeler에 의해 작성된 Microsoft 모델이 저장되는 적절한 데이터베이스를 선택해야 합니다.
- **SQL Server 연결.** 분석 서버에 전달되는 데이터를 저장하기 위해 SQL Server 데이터베이스가 사용하는 DSN 정보를 지정하십시오. Analysis Services 데이터 마이닝 모델 작성을 위해 데이터를 제공하는 데 사용될 ODBC 데이터 소스를 선택하십시오. 플랫 파일 또는 ODBC 데이터 소스에서 제공된 데이터에서 Analysis Services 모델을 작성하는 경우 해당 데이터는 이 ODBC 데이터 소스가 가리키는 SQL Server 데이터베이스에서 작성된 임시 테이블에 자동으로 업로드됩니다.
- **데이터 마이닝 모델을 겹쳐쓰려고 할 때 경고.** 데이터베이스에 저장된 모델이 경고 없이 IBM SPSS Modeler에 의해 겹쳐써지지 않게 하려면 선택하십시오.

참고: 헬퍼 애플리케이션 대화 상자에서 작성된 설정은 다양한 Analysis Services 노드 내부에서 대체될 수 있습니다.

SQL 생성 및 최적화 사용

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

도구 > 스트림 특성 > 옵션

2. 탐색 분할창에서 **최적화** 옵션을 클릭하십시오.

3. **SQL 생성** 옵션이 사용으로 설정되어 있는지 확인하십시오. 이 설정은 데이터베이스 모델링이 작동하기 위해 필수입니다.

4. **SQL 생성 최적화 및 기타 실행 최적화**를 선택하십시오(엄격하게 요구되지 않지만 최적화된 성능을 위해서는 강력하게 권장됨).

## Analysis Services를 사용하여 모델 작성

Analysis Services 모델 작성을 수행하려면 학습 데이터 세트가 SQL Server 데이터베이스 내 테이블 또는 보기에 있어야 합니다. 데이터가 SQL Server에 있지 않거나 IBM SPSS Modeler에서 SQL Server에서 수행할 수 없는 데이터 준비의 일부로 처리되어야 하는 경우 해당 데이터는 모델을 작성하기 전에 SQL Server의 임시 테이블에 자동으로 업로드됩니다.

## Analysis Services 모델 관리

IBM SPSS Modeler를 통해 Analysis Services 모델을 작성하면 IBM SPSS Modeler에서 모델이 작성되고 SQL Server 데이터베이스에서 모델이 작성되거나 바뀝니다. IBM SPSS Modeler 모델은 데이터베이스 서버에 저장된 데이터베이스 모델의 컨텐츠를 참조합니다. IBM SPSS Modeler는 IBM SPSS Modeler 모델과 SQL Server 모델 모두에서 동일한 생성된 모델 키 문자열을 저장하여 일관성 확인을 수행할 수 있습니다.



**MS 의사결정 트리 모델링 노드**는 범주형 속성과 연속형 속성 모두의 예측 모델링에서 사용됩니다. 범주형 속성의 경우 이 노드는 데이터 세트의 입력 열 간 관계를 기반으로 예측을 작성합니다. 예를 들어, 자전거를 구매할 가능성이 있는 고객을 예측하는 시나리오에서 젊은 고객은 10명 중 9명이 자전거를 구입하지만 나이가 많은 고객은 10명 중 2명만 자전거를 구입하는 경우 이 노드는 연령이 자전거 구매의 좋은 예측변수라고 추론합니다. 의사 결정 트리는 특정 결과에 대한 이 경향을 기반으로 예측을 작성합니다. 연속형 속성의 경우 이 알고리즘은 선형 회귀를 사용하여 의사결정 트리가 분할되는 위치를 결정합니다. 둘 이상의 열이 예측 가능으로 설정되거나 입력 데이터에 예측 가능으로 설정된 중첩된 테이블이 포함되어 있는 경우 노드는 각각의 예측 가능한 열에 대해 별도의 의사결정 트리를 작성합니다.



**MS 군집화 모델링 노드**는 반복 기법을 사용하여 데이터 세트의 케이스를 비슷한 특성을 포함하는 군집으로 그룹화합니다. 이 그룹화는 데이터 탐색, 데이터에서 이상 항목 식별 및 예측 작성을 위해 유용합니다. 군집화 모델은 데이터 세트에서 일상적인 관측을 통해 논리적으로 파생시킬 수 없는 관계를 식별합니다. 예를 들어, 자전거로 통근하는 사람들은 일반적으로 직장에서 먼 거리에 살지 않는다고 논리적으로 인식할 수 있습니다. 하지만 이 알고리즘은 명백하지 않은 자전거 통근자에 대한 기타 특성을 찾을 수 있습니다. 군집화 노드는 대상 필드가 지정되지 않은 기타 데이터 마이닝 노드와 다릅니다. 군집화 노드는 데이터가 존재하는 관계와 노드가 식별하는 군집에서 엄격하게 모델을 학습시킵니다.



**MS 연관 규칙 모델링 노드**는 추천 엔진에 유용합니다. 추천 엔진은 고객이 이미 구매했거나 관심을 표시한 항목을 기반으로 고객에게 제품을 추천합니다. 연관 모델은 개별 케이스와 케이스에 포함된 항목 둘 다에 대한 식별자가 포함된 데이터 세트에서 작성됩니다. 케이스의 항목 그룹을 **항목 세트**라고 합니다. 연관 모델은 일련의 항목 세트와 케이스 내에서 해당 항목이 그룹화되는 방식에 대해 설명하는 규칙으로 구성됩니다. 이 알고리즘이 식별하는 규칙은 고객의 장바구니에 이미 있는 항목을 기반으로 고객의 향후 구매 가능성을 예측하는 데 사용할 수 있습니다.



**MS Naive Bayes 모델링 노드**는 대상 필드와 예측변수 필드 간 조건부 확률을 계산하며 열이 독립적이라고 가정합니다. 이 모델은 제안된 모든 예측 변수를 서로 독립된 것으로 취급하므로 *naïve*라는 이름이 사용됩니다. 이 방법은 다른 Analysis Services 알고리즘보다 계산상 덜 집중되므로 모델링의 예비 단계 동안 관계를 신속하게 발견하는 데 유용합니다. 이 노드를 사용하여 데이터의 초기 탐색을 수행한 후 결과를 적용하여 계산 시간이 더 걸리지만 더 정확한 결과를 제공할 수 있는 다른 노드로 추가적인 모델을 작성할 수 있습니다.



**MS 선형 회귀 모델링 노드**는 의사결정 트리 노드의 변형이며 여기서 **MINIMUM\_LEAF\_CASES** 모수는 노드가 마이닝 모델을 학습시키기 위해 사용하는 데이터 세트에 있는 케이스의 총 수 이상으로 설정됩니다. 이 방식으로 모수가 설정되면 노드는 분할을 작성하지 않으므로 선형 회귀를 수행합니다.



**MS 신경망 모델링 노드**는 예측 가능한 속성의 각 상태가 지정된 경우 입력 속성의 가능한 각각의 상태에 대한 확률을 계산한다는 점에서 MS 의사결정 트리 노드와 비슷합니다. 나중에 이 확률을 사용하여 입력 속성을 기반으로 예측된 속성의 결과를 예측할 수 있습니다.



**MS 로지스틱 회귀분석 모델링 노드**는 MS 신경망 노드의 변형이며 여기서 **HIDDEN\_NODE\_RATIO** 모수는 0(영)으로 설정됩니다. 이 설정은 은닉층이 포함되지 않은 신경망 모델을 작성하므로 로지스틱 회귀분석과 동등합니다.



**MS 시계열 모델링 노드**는 시간 경과에 따른 연속 값(예: 제품 판매)의 예측을 위해 최적화된 회귀분석 알고리즘을 제공합니다. 의사결정 트리 등의 다른 Microsoft 알고리즘에는 추세를 예측하기 위해 새 정보의 추가적인 열이 입력으로 필요하지만 시계열 모델에는 필요하지 않습니다. 시계열 모델은 모델을 작성하는 데 사용되는 원래 데이터 세트만 기반으로 추세를 예측할 수 있습니다. 또한 예측을 작성할 때 새 데이터를 모델에 추가하고 추세 분석에서 새 데이터를 자동으로 통합할 수 있습니다. 자세한 정보는 [15 페이지의 『MS 시계열 노드』](#)의 내용을 참조하십시오.



**MS 시퀀스 군집화 모델링 노드**는 데이터에서 정렬된 시퀀스를 식별하고 이 분석의 결과를 군집화 기술과 결합하여 시퀀스 및 기타 속성을 기반으로 군집을 생성합니다. 자세한 정보는 [17 페이지의 『MS 시퀀스 군집화 노드』](#)의 내용을 참조하십시오.

IBM SPSS Modeler 창의 맨 아래에 있는 데이터베이스 모델링 팔레트에서 각각의 노드에 액세스할 수 있습니다.

## 모든 알고리즘 노드에 공통인 설정

다음의 설정은 모든 Analysis Services 알고리즘에 공통입니다.

### 서버 옵션

서버 탭에서는 분석 서버 호스트, 데이터베이스 및 SQL Server 데이터 소스를 구성할 수 있습니다. 여기서 지정된 옵션은 헬퍼 애플리케이션 대화 상자의 Microsoft 탭에서 지정된 옵션을 겹쳐씁니다. 자세한 정보는 [11 페이지의 『Analysis Services와의 통합 사용』](#)의 내용을 참조하십시오.

참고: Analysis Services 모델을 스코어링할 때 이 탭의 변형도 사용할 수 있습니다. 자세한 정보는 [18 페이지의 『Analysis Services 모델 너겟 서버 탭』](#)의 내용을 참조하십시오.

### 모델 옵션

가장 기본적인 모델을 작성하려면 진행하기 전에 모델 탭에서 옵션을 지정해야 합니다. 스코어링 방법 및 기타 고급 옵션을 고급 탭에서 사용할 수 있습니다.

다음과 같은 기본 모델링 옵션을 사용할 수 있습니다.

**모델 이름.** 노드가 실행될 때 작성되는 모델에 지정된 이름을 지정합니다.

- 자동.** 목표 또는 ID 필드 이름이나 목표가 지정되지 않은 경우(예: 군집 모델) 모델 유형 이름에 따라 모델 이름을 자동으로 생성합니다.
- 사용자 정의.** 작성된 모델에 대한 사용자 정의 이름을 지정할 수 있게 합니다.

**파티션된 데이터 사용.** 현재 파티션 필드를 기반으로 학습, 검정 및 검증을 위한 개별 서브세트 또는 표본으로 데이터를 분할합니다. 하나의 표본을 사용하여 모델을 작성하고 개별 표본을 사용하여 해당 모델을 검정하면 현재 데이터와 비슷한 더 큰 데이터 세트로 모델이 일반화되는 정도를 표시할 수 있습니다. 스트림에 지정된 파티션 필드가 없으면 이 옵션은 무시됩니다.

**드릴스루 사용.** 표시된 경우 이 옵션을 사용하면 모델을 쿼리하여 모델에 포함된 케이스에 대한 세부사항을 학습할 수 있습니다.

**고유 필드.** 드롭 다운 목록에서 각각의 케이스를 고유하게 식별하는 필드를 선택하십시오. 일반적으로 이 필드는 ID 필드(예: **CustomerID**)입니다.

## MS 의사결정 트리 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 군집화 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS Naive Bayes 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 선형 회귀 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 신경망 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 로지스틱 회귀분석 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 연관 규칙 노드

MS 연관 규칙 모델링 노드는 추천 엔진에 유용합니다. 추천 엔진은 고객이 이미 구매했거나 관심을 표시한 항목을 기반으로 고객에게 제품을 추천합니다. 연관 모델은 개별 케이스와 케이스에 포함된 항목 둘 다에 대한 식별자가 포함된 데이터 세트에서 작성됩니다. 케이스의 항목 그룹을 **항목 세트**라고 합니다.

연관 모델은 일련의 항목 세트와 케이스 내에서 해당 항목이 그룹화되는 방식에 대해 설명하는 규칙으로 구성됩니다. 이 알고리즘이 식별하는 규칙은 고객의 장바구니에 이미 있는 항목을 기반으로 고객의 향후 구매 가능성을 예측하는 데 사용할 수 있습니다.

표 형식 데이터의 경우 이 알고리즘은 각각의 생성된 추천(\$M-field)에 대한 확률(\$MP-field)을 나타내는 스코어를 작성합니다. 트랜잭션 형식 데이터의 경우 각각의 생성된 추천(\$M-field)에 대해 지원(\$MS-field), 확률(\$MP-field) 및 조정된 확률(\$MAP-field)에 대한 스코어가 작성됩니다.

### 요구사항

트랜잭션 연관 모델에 대한 요구사항은 다음과 같습니다.

- **고유 필드.** 연관 규칙 모델에는 레코드를 고유하게 식별하는 키가 필요합니다.
- **ID 필드.** 트랜잭션 형식 데이터를 사용하여 MS 연관 규칙 모델을 작성하는 경우에는 각 트랜잭션을 식별하는 ID 필드가 필요합니다. ID 필드는 고유 필드와 동일하게 설정할 수 있습니다.
- **하나 이상의 입력 필드.** 연관 규칙 알고리즘에는 하나 이상의 입력 필드가 필요합니다.
- **대상 필드.** 트랜잭션 데이터를 사용하여 MS 연관 모델을 작성하는 경우에는 대상 필드가 트랜잭션 필드여야 합니다(예: 사용자가 구입한 제품).

## MS 연관 규칙 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## MS 시계열 노드

MS 시계열 모델링 노드는 두 가지 유형의 예측을 지원합니다.

- 미래
- 히스토리

미래 예측은 히스토리 데이터의 끝을 넘어선 지정된 수의 기간에 대한 대상 필드 값을 추정하며 항상 수행됩니다. 히스토리 예측은 히스토리 데이터에 실제 값을 가진 지정된 수의 기간에 대한 대상 필드 값을 추정합니다. 히스토

리 예측을 통해 실제 히스토리 값과 예측값을 비교하여 모델의 품질을 평가할 수 있습니다. 예측 시작점의 값은 히스토리 예측이 수행되는지 여부를 판별합니다.

IBM SPSS Modeler 시계열 노드와는 달리 MS 시계열 노드에는 선행 시간 간격 노드가 필요하지 않습니다. 추가적인 차이점은 기본적으로 시계열 데이터의 모든 히스토리 행이 아니라 예측된 행에 대해서만 스코어가 생성된다는 점입니다.

### 요구사항

MS 시계열 모델에 대한 요구사항은 다음과 같습니다.

- **단일 키 시간 필드.** 각각의 모델에는 모델이 사용할 시간 조각을 정의하는 케이스 시리즈로 사용되는 하나의 숫자 또는 날짜 필드가 포함되어 있어야 합니다. 키 시간 필드에 대한 데이터 유형은 날짜/시간 데이터 유형 또는 숫자 데이터 유형일 수 있습니다. 하지만 이 필드에는 연속형 값이 포함되어 있어야 하며 값은 각각의 시리즈에 대해 고유해야 합니다.
- **단일 대상 필드.** 각각의 모델에서 하나의 대상 필드만 지정할 수 있습니다. 대상 필드의 데이터 유형은 연속형 값을 가져야 합니다. 예를 들어, 수입, 판매 또는 온도 등의 숫자 속성이 시간 경과에 따라 어떻게 변하는지 예측할 수 있습니다. 하지만 범주형 값(예: 구매 상태 또는 교육 수준)을 대상 필드로 포함하는 필드는 사용할 수 없습니다.
- **하나 이상의 입력 필드.** MS 시계열 알고리즘에는 하나 이상의 입력 필드가 필요합니다. 입력 필드의 데이터 유형은 연속형 값을 가져야 합니다. 비연속형 입력 필드는 모델 작성 시 무시됩니다.
- **데이터 세트가 정렬되어야 함.** 입력 데이터 세트는 키 시간 필드에서 정렬되어야 합니다. 그렇지 않으면 모델 작성성이 중단되고 오류가 생성됩니다.

## MS 시계열 모델 옵션

**모델 이름.** 노드가 실행될 때 작성되는 모델에 지정된 이름을 지정합니다.

- **자동.** 목표 또는 ID 필드 이름이나 목표가 지정되지 않은 경우(예: 군집 모델) 모델 유형 이름에 따라 모델 이름을 자동으로 생성합니다.
- **사용자 정의.** 작성된 모델에 대한 사용자 정의 이름을 지정할 수 있게 합니다.

**파티션된 데이터 사용.** 파티션 필드가 정의된 경우 이 옵션을 사용하면 학습 파티션의 데이터만 사용하여 모델을 작성합니다.

**드릴스루 사용.** 표시된 경우 이 옵션을 사용하면 모델을 쿼리하여 모델에 포함된 케이스에 대한 세부사항을 학습할 수 있습니다.

**고유 필드.** 드롭 다운 목록에서 시계열 모델을 작성하는 데 사용되는 키 시간 필드를 선택하십시오.

## MS 시계열 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

히스토리 예측을 작성하는 경우 스코어링 결과에 포함할 수 있는 히스토리 단계의 수는 (HISTORIC\_MODEL\_COUNT \* HISTORIC\_MODEL\_GAP)의 값에 의해 결정됩니다. 기본적으로 이 제한은 10이며 이는 10개의 히스토리 예측만 작성됨을 의미합니다. 예를 들어, 이 경우 모델 너깃의 설정 탭에서 히스토리 예측에 대해 -10 미만의 값을 입력하면 오류가 발생합니다([19 페이지의 『MS 시계열 모델 너깃 설정 탭』](#) 참조). 더 많은 히스토리 예측을 보려는 경우 HISTORIC\_MODEL\_COUNT 또는 HISTORIC\_MODEL\_GAP의 값을 늘릴 수 있습니다. 하지만 이를 수행하면 모델의 작성 시간이 증가합니다.

## MS 시계열 설정 옵션

**추정 시작.** 예측을 시작할 기간을 지정하십시오.

- **시작 위치: 새 예측.** 향후 예측을 시작할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 오프셋으로 표시됨). 예를 들어, 히스토리 데이터가 12/99에서 종료된 경우 01/00에서 예측을 시작하려면 값으로 1을 사용합니다. 하지만 03/00에서 예측을 시작하려면 값으로 3을 사용합니다.

- **시작 위치: 히스토리 예측.** 히스토리 예측을 시작할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 음수 오프셋으로 표현됨). 예를 들어, 히스토리 데이터가 12/99에서 종료된 경우 데이터의 마지막 5개 기간에 대한 히스토리 예측을 작성하려면 값으로 -5를 사용합니다.

**추정 종료.** 예측을 중지할 기간을 지정하십시오.

- **예측 단계 종료.** 예측을 중지할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 오프셋으로 표현됨). 예를 들어, 히스토리 데이터가 12/99에서 종료되는 경우 6/00에서 예측을 중지하려면 여기서 값으로 6을 사용합니다. 향후 예측의 경우 같은 항상 **시작 위치** 값 이상이어야 합니다.

## MS 시퀀스 군집화 노드

MS 시퀀스 군집화 노드는 다음과 같은 경로(또는 시퀀스)를 사용하여 링크할 수 있는 이벤트가 포함된 데이터를 탐색하는 시퀀스 분석 알고리즘을 사용합니다. 이에 대한 예제로는 사용자가 웹 사이트를 탐색하거나 찾아볼 때 작성된 클릭 경로, 온라인 소매업체에서 고객이 장바구니에 항목을 추가하는 순서 등이 있습니다. 이 알고리즘은 동일한 시퀀스를 그룹화(또는 군집화)하여 가장 일반적인 시퀀스를 찾습니다.

### 요구사항

Microsoft 시퀀스 군집화 모델에 대한 요구사항은 다음과 같습니다.

- **ID 필드.** Microsoft 시퀀스 군집화 알고리즘을 사용하려면 시퀀스 정보를 트랜잭션 형식으로 저장해야 합니다. 이를 위해 각 트랜잭션을 식별하는 ID 필드가 필요합니다.
- **하나 이상의 입력 필드.** 이 알고리즘에는 하나 이상의 입력 필드가 필요합니다.
- **시퀀스 필드.** 이 알고리즘에는 연속형 측정 수준을 가져야 하는 시퀀스 식별자 필드도 필요합니다. 예를 들어, 필드가 시퀀스에서 이벤트를 식별하는 경우 웹 페이지 식별자, 정수 또는 텍스트 문자열을 사용할 수 있습니다. 각 시퀀스에 대해 하나의 시퀀스 식별자만 허용되며 각 모델에서는 한 유형의 시퀀스만 허용됩니다. 시퀀스 필드는 ID 및 고유 필드와 달라야 합니다.
- **대상 필드.** 대상 필드는 시퀀스 군집화 모델 작성 시 필수입니다.
- **고유 필드.** 시퀀스 군집화 모델에는 레코드를 고유하게 식별하는 키 필드가 필요합니다. ID 필드와 동일하도록 고유 필드를 설정할 수 있습니다.

## MS 시퀀스 군집화 필드 옵션

모든 모델링 노드에는 필드 탭이 있으며 이 탭에서는 모델 작성 시 사용할 필드를 지정합니다.

시퀀스 군집화 모델을 작성하려면 먼저 목표 및 입력으로 사용할 필드를 지정해야 합니다. MS 시퀀스 군집화 노드의 경우 업스트림 유형 노드의 필드 정보는 사용할 수 없으므로 여기서 필드 설정을 지정해야 합니다.

**ID.** 목록에서 ID 필드를 선택합니다. 숫자 또는 기호 필드를 ID 필드로 사용할 수 있습니다. 이 필드의 각 고유한 값은 분석의 특정 단위를 표시해야 합니다. 예를 들어, 장바구니 애플리케이션에서 각 ID는 단일 고객을 나타낼 수 있습니다. 웹 로그 분석 애플리케이션의 경우 각 ID는 컴퓨터(IP 주소별) 또는 사용자(로그인 데이터별)를 나타낼 수 있습니다.

**입력.** 모델에 대한 입력 필드를 선택하십시오. 이 필드는 시퀀스 모델링에 관심 있는 이벤트가 포함된 필드입니다.

**시퀀스.** 시퀀스 식별자 필드로 사용할 필드를 목록에서 선택하십시오. 예를 들어, 필드가 시퀀스에서 이벤트를 식별하는 경우 웹 페이지 식별자, 정수 또는 텍스트 문자열을 사용할 수 있습니다. 각 시퀀스에 대해 하나의 시퀀스 식별자만 허용되며 각 모델에서는 한 유형의 시퀀스만 허용됩니다. 시퀀스 필드는 ID 필드(이 탭에서 지정됨) 및 고유 필드(모델 탭에서 지정됨)와 달라야 합니다.

**목표.** 대상 필드(즉, 시퀀스 데이터를 기반으로 값을 예측하는 필드)로 사용할 필드를 선택하십시오.

## MS 시퀀스 군집화 고급 옵션

고급 탭에서 사용할 수 있는 옵션은 선택된 스트림의 구조에 따라 다를 수 있습니다. 선택된 Analysis Services 모델 노드에 대한 고급 옵션과 관련된 전체 세부사항은 사용자 인터페이스 필드 수준 도움말을 참조하십시오.

## Analysis Services 모델 스코어링

모델 스코어링은 SQL Server에서 발생하며 Analysis Services에 의해 수행됩니다. 데이터를 IBM SPSS Modeler에서 제공하거나 IBM SPSS Modeler에서 준비해야 하는 경우에는 데이터 세트를 임시 테이블에 업로드해야 할 수 있습니다. In-Database 마이닝을 사용하여 IBM SPSS Modeler에서 작성하는 모델은 실제로 원격 데이터 마이닝 또는 데이터베이스 서버에서 보유되는 원격 모델입니다. 이는 Microsoft Analysis Services 알고리즘을 사용하여 작성된 모델을 찾아서 스코어링할 때 이해할 중요한 차이입니다.

IBM SPSS Modeler에서는 일반적으로 하나의 예측 및 연관된 확률 또는 신뢰도가 전달됩니다.

모델 스코어링 예제는 [20 페이지의 『Analysis Services 마이닝 예제』](#)의 내용을 참조하십시오.

## 모든 Analysis Services 모델에 공통인 설정

다음의 설정은 모든 Analysis Services 모델에 공통입니다.

### Analysis Services 모델 너깃 서버 탭

서버 탭은 In-Database 마이닝에 대한 연결을 지정하는 데 사용됩니다. 이 탭은 고유 모델 키도 제공합니다. 이 키는 모델이 작성될 때 무작위로 생성되며 IBM SPSS Modeler의 모델과 Analysis Services 데이터베이스에 저장된 모델 오브젝트의 설명에도 저장됩니다.

서버 탭에서는 스코어링 조작을 위해 SQL Server 데이터 소스와 분석 서버 호스트 및 데이터베이스를 구성할 수 있습니다. 여기서 지정된 옵션은 IBM SPSS Modeler의 헬퍼 애플리케이션 또는 모델 작성 대화 상자에서 지정된 옵션을 겹쳐씁니다. 자세한 정보는 [11 페이지의 『Analysis Services와의 통합 사용』](#)의 내용을 참조하십시오.

**모델 GUID.** 모델 키가 여기에 표시됩니다. 이 키는 모델이 작성될 때 무작위로 생성되며 IBM SPSS Modeler의 모델과 Analysis Services 데이터베이스에 저장된 모델 오브젝트의 설명에도 저장됩니다.

**검사.** Analysis Services 데이터베이스에 저장된 모델의 키에 대해 모델 키를 확인하려면 이 단추를 클릭하십시오. 이는 모델이 분석 서버에 여전히 존재하는지 확인할 수 있게 하며 모델의 구조가 변경되지 않았음을 나타냅니다.

**참고:** 확인 단추는 스코어링에 대비해 스트림 캔버스에 추가된 모델의 경우에만 사용할 수 있습니다. 확인에 실패하는 경우에는 모델이 삭제되었거나 서버의 다른 모델로 바뀌었는지 조사하십시오.

**보기.** 의사결정 트리 모형의 그래픽 보기是为了 클릭하십시오. 의사결정 트리 뷰어는 IBM SPSS Modeler의 기타 의사결정 트리 알고리즘에 의해 공유되며 기능은 동일합니다.

### Analysis Services 모델 너깃 요약 탭

모델 너깃의 요약 탭에서는 모델 자체(분석), 모델에 사용된 필드(필드), 모델 작성 시 사용된 설정(작성 설정), 모델 학습(학습 요약)에 대한 정보를 표시합니다.

먼저 노드를 찾아볼 때 요약 탭 결과를 접습니다. 관심이 있는 결과를 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 펼치거나 모두 펼치기 단추를 클릭하여 모든 결과를 표시합니다. 결과 보기를 완료한 경우 결과를 숨기려면 펼치기 제어를 사용하여 숨기려는 특정 결과를 접거나 모두 접기 단추를 클릭하여 모든 결과를 접으십시오.

**분석.** 특정 모델에 대한 정보를 표시합니다. 이 모델 너깃에 연결된 분석 노드를 실행한 경우에는 해당 분석의 정보도 이 섹션에 표시됩니다.

**필드.** 모델을 작성할 때 목표로 사용되는 필드와 입력을 나열합니다.

**작성 설정.** 모델을 작성할 때 사용되는 설정에 대한 정보를 포함합니다.

**학습 요약.** 모델 유형, 이를 작성하는데 사용된 스트림, 이를 작성한 사용자, 작성 시점, 모델 작성 시 경과 시간을 표시합니다.

## MS 시계열 모델 너깃

MS 시계열 모델은 히스토리 데이터가 아니라 예측된 기간에 대한 스코어만 생성합니다.

다음 표에는 모델에 추가되는 필드가 표시됩니다.

표 1. 모델에 추가되는 필드	
필드 이름	설명
\$M-필드	필드의 예측값입니다.
\$Var-필드	필드의 계산된 분산
\$Stdev-필드	필드의 표준 편차

## MS 시계열 모델 너깃 서버 탭

서버 탭은 In-Database 마이닝에 대한 연결을 지정하는 데 사용됩니다. 이 탭은 고유 모델 키도 제공합니다. 이 키는 모델이 작성될 때 무작위로 생성되며 IBM SPSS Modeler의 모델과 Analysis Services 데이터베이스에 저장된 모델 오브젝트의 설명에도 저장됩니다.

서버 탭에서는 스코어링 조작을 위해 SQL Server 데이터 소스와 분석 서버 호스트 및 데이터베이스를 구성할 수 있습니다. 여기서 지정된 옵션은 IBM SPSS Modeler의 헬퍼 애플리케이션 또는 모델 작성 대화 상자에서 지정된 옵션을 겹쳐씁니다. 자세한 정보는 [11 페이지의 『Analysis Services와의 통합 사용』](#)의 내용을 참조하십시오.

**모델 GUID.** 모델 키가 여기에 표시됩니다. 이 키는 모델이 작성될 때 무작위로 생성되며 IBM SPSS Modeler의 모델과 Analysis Services 데이터베이스에 저장된 모델 오브젝트의 설명에도 저장됩니다.

**검사.** Analysis Services 데이터베이스에 저장된 모델의 키에 대해 모델 키를 확인하려면 이 단추를 클릭하십시오. 이는 모델이 분석 서버에 여전히 존재하는지 확인할 수 있게 하며 모델의 구조가 변경되지 않았음을 나타냅니다.

**참고:** 확인 단추는 스코어링에 대비해 스트림 캔버스에 추가된 모델의 경우에만 사용할 수 있습니다. 확인에 실패하는 경우에는 모델이 삭제되었거나 서버의 다른 모델로 바뀌었는지 조사하십시오.

**보기.** 시계열 모델의 그래픽 보기是为了 클릭하십시오. Analysis Services는 완료된 모델을 트리로 표시합니다. 예측된 미래 값과 함께 시간 경과에 따른 대상 필드의 히스토리 값을 표시하는 그래프도 볼 수 있습니다.

자세한 정보는 MSDN 라이브러리에서 시계열 뷰어에 대한 설명을 참조하십시오(<http://msdn.microsoft.com/en-us/library/ms175331.aspx>).

## MS 시계열 모델 너깃 설정 탭

**추정 시작.** 예측을 시작할 기간을 지정하십시오.

- 시작 위치: 새 예측.** 향후 예측을 시작할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 오프셋으로 표현됨). 예를 들어, 히스토리 데이터가 12/99에서 종료된 경우 01/00에서 예측을 시작하려면 값으로 1을 사용합니다. 하지만 03/00에서 예측을 시작하려면 값으로 3을 사용합니다.
- 시작 위치: 히스토리 예측.** 히스토리 예측을 시작할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 음수 오프셋으로 표현됨). 예를 들어, 히스토리 데이터가 12/99에서 종료된 경우 데이터의 마지막 5개 기간에 대한 히스토리 예측을 작성하려면 값으로 -5를 사용합니다.

**추정 종료.** 예측을 중지할 기간을 지정하십시오.

- 예측 단계 종료.** 예측을 중지할 기간입니다(히스토리 데이터의 마지막 기간으로부터의 오프셋으로 표현됨). 예를 들어, 히스토리 데이터가 12/99에서 종료되는 경우 6/00에서 예측을 중지하려면 여기서 값으로 6을 사용합니다. 향후 예측의 경우 값은 항상 시작 위치 값 이상이어야 합니다.

## MS 시퀀스 군집화 모델 너깃

다음 표에는 MS 시퀀스 군집화 모델에 추가되는 필드가 표시됩니다(여기서 필드는 대상 필드의 이름임).

표 2. 모델에 추가되는 필드	
필드 이름	설명
\$MC-필드	이 시퀀스가 속하는 군집에 대한 예측입니다.
\$MCP-필드	이 시퀀스가 예측된 군집에 속하는 확률입니다.

표 2. 모델에 추가되는 필드 (계속)	
필드 이름	설명
\$MS-필드	필드의 예측값입니다.
\$MSP-필드	\$MS-필드 값이 올바를 확률입니다.

## 모델 내보내기 및 노드 생성

모델 요약 및 구조를 텍스트 및 HTML 형식 파일로 내보낼 수 있습니다. 해당되는 경우 적절한 선택 및 필터 노드를 생성할 수 있습니다.

IBM SPSS Modeler의 기타 모델 너깃과 마찬가지로 Microsoft Analysis Services 모델 너깃은 레코드 및 필드 조작 노드의 직접 생성을 지원합니다. 모델 너깃 메뉴 생성 옵션을 사용하면 다음과 같은 노드를 생성할 수 있습니다.

- 선택 노드(모델 탭에서 항목이 선택되는 경우에만)
- 필터 노드

## Analysis Services 마이닝 예제

IBM SPSS Modeler와 함께 MS Analysis Services 데이터 마이닝을 사용하는 것을 보여주는 다수의 샘플 스트림이 포함되어 있습니다. 이 스트림은 다음 위치에 있는 IBM SPSS Modeler 설치 폴더에서 찾을 수 있습니다.

|Demos|Database\_Modelling|Microsoft

참고: Demos 폴더는 Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 액세스할 수 있습니다.

## 예제 스트림: 의사결정 트리

다음의 스트림은 MS Analysis Services에서 제공하는 의사결정 트리 알고리즘을 사용하여 데이터베이스 마이닝 프로세스의 예로 순차적으로 함께 사용될 수 있습니다.

표 3. 의사결정 트리 - 예제 스트림	
스트림	설명
1_upload_data.str	플랫 파일에서 데이터베이스로 데이터를 정리하고 업로드하는 데 사용됩니다.
2_explore_data.str	IBM SPSS Modeler를 사용한 데이터 탐색의 예를 제공합니다.
3_build_model.str	데이터베이스의 원래 알고리즘을 사용하여 모델을 작성합니다.
4_evaluate_model.str	IBM SPSS Modeler를 사용하여 모형을 평가하는 예로 사용됩니다.
5_deploy_model.str	In-Database 스코어링에 대한 모델을 배포합니다.

참고: 예를 실행하려면 스트림이 순서대로 실행되어야 합니다. 또한 사용할 데이터베이스에 대한 유효한 데이터 소스를 참조하기 위해 각 스트림의 소스 및 모델링 노드가 업데이트되어야 합니다.

예제 스트림에서 사용된 데이터 세트는 신용카드 애플리케이션과 관련되며 분류 문제점에 범주형 예측변수와 연속형 예측변수의 혼합을 제공합니다. 이 데이터 세트에 대한 자세한 정보는 샘플 스트림과 동일한 폴더의 crx.names 파일을 참조하십시오.

이 데이터 세트는 UCI Machine Learning Repository <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>에서 사용 가능합니다.

## 예제 스트림: 데이터 업로드

첫 번째 예제 스트림인 *1\_upload\_data.str*은 플랫 파일의 데이터를 정리하고 SQL Server로 업로드하는 데 사용됩니다.

Analysis Services 데이터 마이닝에는 키 필드가 필요하므로 이 초기 스트림에서는 파생 노드를 사용하여 IBM SPSS Modeler @INDEX 함수로 고유 값 1,2,3을 가진 KEY라는 새 필드를 데이터 세트에 추가합니다.

후속 채움 노드는 결측값 처리에 사용되며 텍스트 파일 *crx.data*에서 읽어온 비어 있는 필드를 *NULL* 값으로 바꿉니다.

## 예제 스트림: 데이터 탐색

두 번째 예제 스트림인 *2\_explore\_data.str*은 요약 통계 및 그래프를 포함하여 데이터의 일반적인 개요를 얻기 위해 데이터 검토 노드를 사용하는 방법을 보여줍니다.

데이터 검토 보고서에서 그래프를 두 번 클릭하면 지정된 필드를 더 깊게 탐색할 수 있도록 자세한 그래프가 생성됩니다.

## 예제 스트림: 모델 작성

세 번째 예제 스트림인 *3\_build\_model.str*은 IBM SPSS Modeler에서 모델 작성을 보여줍니다. 데이터베이스 모델을 스트림에 연결한 후 두 번 클릭하여 작성 설정을 지정할 수 있습니다.

대화 상자의 모델 탭에서는 다음을 지정할 수 있습니다.

1. 키 필드를 고유 ID 필드로 선택하십시오.

고급 탭에서는 모델 작성을 위한 설정을 미세 조정할 수 있습니다.

실행하기 전에 모델 작성을 위해 올바른 데이터베이스를 지정했는지 확인하십시오. 서버 탭을 사용하여 설정을 조정하십시오.

## 예제 스트림: 모델 평가

네 번째 예제 스트림인 *4\_evaluate\_model.str*은 In-Database 모델링에 대해 IBM SPSS Modeler를 사용할 때의 장점을 보여줍니다. 모델을 실행하고 나면 해당 모델을 다시 데이터 스트림에 추가하고 IBM SPSS Modeler에서 제공된 여러 도구를 사용하여 해당 모델을 평가할 수 있습니다.

### 모델링 결과 보기

모델 너깃을 두 번 클릭하여 결과를 탐색할 수 있습니다. 요약 탭은 결과의 규칙-트리 보기와 의사결정 트리 모형의 그래픽 보기에 대한 보기 단추(서버 탭에 있음)를 클릭할 수도 있습니다.

### 모델 결과 평가

샘플 스트림의 분석 노드는 각 예측 필드 및 해당 대상 필드 간 일치 패턴을 보여주는 일치 교차표를 작성합니다. 분석 노드를 실행하여 결과를 확인하십시오.

샘플 스트림의 평가 노드는 모델에 의해 작성된 정확도 개선사항을 표시하도록 설계된 Gains 차트를 작성할 수 있습니다. 평가 노드를 실행하여 결과를 확인하십시오.

## 예제 스트림: 모델 배포

모델의 정확도에 만족한 경우에는 외부 애플리케이션과 함께 사용하거나 데이터베이스에 다시 게시하기 위해 해당 모델을 배포할 수 있습니다. 최종 예제 스트림인 *5\_deploy\_model.str*에서는 데이터를 CREDIT 테이블에서 읽어온 후 데이터베이스 내보내기 노드를 사용하여 스코어링하여 CREDITSCORES 테이블에 게시합니다.

스트림을 실행하면 다음의 SQL이 생성됩니다.

```
DROP TABLE CREDITSCORES  
CREATE TABLE CREDITSCORES ( "field1" varchar(1), "field2" varchar(255), "field3" float, "field4" varchar(1), "field5" varchar(2), "field6" varchar(2), "field7" varchar(2), "field8" float, "field9" varchar(1), "field10" varchar(1), "field11" int, "field12" varchar(1), "field13" varchar(1), "field14" int, "field15" int, "field16" varchar(1), "KEY" int, "$M-field16" varchar(9), "$MC-field16" float )  
INSERT INTO CREDITSCORES ("field1", "field2", "field3", "field4", "field5", "field6", "field7", "field8",
```

```

"field9","field10","field11","field12","field13","field14","field15","field16",
"KEY","$M-field16","$MC-field16")
SELECT T0.C0 AS C0,T0.C1 AS C1,T0.C2 AS C2,T0.C3 AS C3,T0.C4 AS C4,T0.C5 AS C5,
T0.C6 AS C6,T0.C7 AS C7,T0.C8 AS C8,T0.C9 AS C9,T0.C10 AS C10,
T0.C11 AS C11,T0.C12 AS C12,T0.C13 AS C13,T0.C14 AS C14,
T0.C15 AS C15,T0.C16 AS C16,T0.C17 AS C17,T0.C18 AS C18
FROM (
    SELECT CONVERT(NVARCHAR,[TA].[field1]) AS C0, CONVERT(NVARCHAR,[TA].[field2]) AS C1,
    [TA].[field3] AS C2, CONVERT(NVARCHAR,[TA].[field4]) AS C3,
    CONVERT(NVARCHAR,[TA].[field5]) AS C4, CONVERT(NVARCHAR,[TA].[field6]) AS C5,
    CONVERT(NVARCHAR,[TA].[field7]) AS C6, [TA].[field8] AS C7,
    CONVERT(NVARCHAR,[TA].[field9]) AS C8, CONVERT(NVARCHAR,[TA].[field10]) AS C9,
    [TA].[field11] AS C10, CONVERT(NVARCHAR,[TA].[field12]) AS C11,
    CONVERT(NVARCHAR,[TA].[field13]) AS C12, [TA].[field14] AS C13,
    [TA].[field15] AS C14, CONVERT(NVARCHAR,[TA].[field16]) AS C15,
    [TA].[KEY] AS C16, CONVERT(NVARCHAR,[TA].[$M-field16]) AS C17,
    [TA].[$MC-field16] AS C18
FROM openrowset('MSOLAP',
'Datasource=localhost;Initial catalog=FoodMart 2000',
'SELECT [T].[C0] AS [field1],[T].[C1] AS [field2],[T].[C2] AS [field3],
[T].[C3] AS [field4],[T].[C4] AS [field5],[T].[C5] AS [field6],
[T].[C6] AS [field7],[T].[C7] AS [field8],[T].[C8] AS [field9],
[T].[C9] AS [field10],[T].[C10] AS [field11],[T].[C11] AS [field12],
[T].[C12] AS [field13],[T].[C13] AS [field14],[T].[C14] AS [field15],
[T].[C15] AS [field16],[T].[C16] AS [KEY],[CREDIT1].[field16] AS [$M-field16],
PredictProbability([CREDIT1].[field16]) AS [$MC-field16]
FROM [CREDIT1] PREDICTION JOIN
openrowset('MSDASQL',
'Dsn=LocalServer;Uid=''', ''SELECT T0."field1" AS C0,T0."field2" AS C1,
T0."field3" AS C2,T0."field4" AS C3,T0."field5" AS C4,T0."field6" AS C5,
T0."field7" AS C6,T0."field8" AS C7,T0."field9" AS C8,T0."field10" AS C9,
T0."field11" AS C10,T0."field12" AS C11,T0."field13" AS C12,
T0."field14" AS C13,T0."field15" AS C14,T0."field16" AS C15,
T0."KEY" AS C16 FROM "dbo".CREDITDATA T0') AS [T]
ON [T].[C2] = [CREDIT1].[field3] and [T].[C7] = [CREDIT1].[field8]
and [T].[C8] = [CREDIT1].[field9] and [T].[C9] = [CREDIT1].[field10]
and [T].[C10] = [CREDIT1].[field11] and [T].[C11] = [CREDIT1].[field12]
and [T].[C14] = [CREDIT1].[field15]') AS [TA]
) T0

```

# 제 4 장 Oracle Data Mining을 사용한 데이터베이스 모델링

## Oracle Data Mining 정보

IBM SPSS Modeler는 Oracle RDBMS 내에서 단단하게 임베드된 데이터 마이닝 알고리즘 패밀리를 제공하는 Oracle 데이터 마이닝(ODM)과의 통합을 지원합니다. 이 기능은 IBM SPSS Modeler 그래픽 사용자 인터페이스 및 워크플로우 지향 개발 환경을 통해 액세스할 수 있으며 이를 통해 고객이 ODM이 제공하는 데이터 마이닝 알고리즘을 사용할 수 있습니다.

IBM SPSS Modeler는 Oracle 데이터 마이닝에서 제공하는 다음 알고리즘의 통합을 지원합니다.

- Naive Bayes
- 적응형 베이스
- SVM(Support Vector Machine)
- GLM(Generalized Linear Model)\*
- 의사결정 트리
- O-군집
- K-평균
- NMF(Nonnegative Matrix Factorization)
- Apriori
- MDL(Minimum Descriptor Length)
- 속성 중요도(AI)

\* 11g R1 전용

## Oracle과의 통합을 위한 요구사항

Oracle Data Mining을 사용하여 In-Database 모델링을 수행하기 위해서는 다음과 같은 조건이 충족되어야 합니다. 데이터베이스 관리자에게 문의하여 이 조건이 충족되는지 확인해야 할 수 있습니다.

- Windows 또는 UNIX에서 IBM SPSS Modeler Server 설치에 대해 또는 로컬 모드로 실행 중인 IBM SPSS Modeler
- Oracle Data Mining 옵션이 설치된 Oracle 10gR2 또는 11gR1(10.2 데이터베이스 이상)

**참고:** 10gR2는 일반화 선형 모형(11gR1이 필요함)을 제외한 모든 데이터베이스 모델링 알고리즘에 대한 지원을 제공합니다.

- 아래에 설명된 대로 Oracle에 연결하는 데 필요한 ODBC 데이터 소스.

**참고:** 데이터베이스 모델링 및 SQL 최적화를 위해서는 IBM SPSS Modeler 컴퓨터에서 IBM SPSS Modeler Server 연결이 사용으로 설정되어 있어야 합니다. 이 설정이 사용 가능하면 데이터베이스 알고리즘에 액세스하고, SQL을 IBM SPSS Modeler에서 직접 푸시백하고, IBM SPSS Modeler Server에 액세스할 수 있습니다. 현재 라이센스 상태를 검증하려면 IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

### 도움말 > 정보 > 추가 세부사항

연결성이 사용 가능으로 설정되면 라이센스 상태 탭에 서버 사용 가능 옵션이 표시됩니다.

## Oracle과의 통합 사용

Oracle Data Mining과의 IBM SPSS Modeler 통합을 사용하려면 Oracle을 구성하고 ODBC 소스를 작성하고 IBM SPSS Modeler 헬퍼 애플리케이션에서 통합을 사용으로 설정하고 SQL 생성 및 최적화를 사용으로 설정해야 합니다.

### Oracle 구성

Oracle Data Mining을 설치하고 구성하려면 Oracle 문서(특히 *Oracle 관리자 안내서*)에서 자세한 내용을 참조하십시오.

### Oracle에 대한 ODBC 소스 작성

Oracle과 IBM SPSS Modeler 간 연결을 사용하려면 ODBC 시스템 데이터 소스 이름(DSN)을 작성해야 합니다.

DSN을 작성하기 전에 ODBC 데이터 소스 및 드라이버와 IBM SPSS Modeler에서의 데이터베이스 지원에 대한 기본적인 이해가 필요합니다.

IBM SPSS Modeler Server에 대해 분산 모드에서 실행 중인 경우에는 서버 컴퓨터에서 DSN을 작성하십시오. 로컬(클라이언트) 모드에서 실행 중인 경우에는 클라이언트 컴퓨터에서 DSN을 작성하십시오.

1. ODBC 드라이버를 설치하십시오. 이 드라이버는 이 릴리스와 함께 제공된 IBM SPSS Data Access Pack 설치 디스크에 있습니다. *setup.exe* 파일을 실행하여 설치 프로그램을 시작하고 모든 관련 드라이버를 선택하십시오. 화면에 표시되는 지시사항에 따라 드라이버를 설치하십시오.

- a. DSN 작성.

참고: 메뉴 순서는 Windows 버전에 따라 다릅니다.

- **Windows XP.** 시작 메뉴에서 제어판을 선택하십시오. 관리 도구를 두 번 클릭한 후 데이터 소스(**ODBC**)를 두 번 클릭하십시오.
- **Windows Vista.** 시작 메뉴에서 제어판을 선택한 후 시스템 유지보수를 선택하십시오. 관리 도구를 두 번 클릭하고 데이터 소스(**ODBC**)를 선택한 후 열기를 클릭하십시오.
- **Windows 7.** 시작 메뉴에서 제어판, 시스템 & 보안, 관리 도구를 차례로 선택하십시오. 데이터 소스 (**ODBC**)를 선택한 후 열기를 클릭하십시오.

- b. 시스템 DSN 탭으로 이동한 후 추가를 클릭하십시오.

2. **SPSS OEM 6.0 Oracle Wire Protocol** 드라이버를 선택하십시오.

3. 완료를 클릭하십시오.

4. ODBC Oracle Wire Protocol 드라이버 설정 화면에서 선택한 데이터 소스 이름, Oracle 서버의 호스트 이름, 연결의 포트 번호 및 사용 중인 Oracle 인스턴스의 SID를 입력하십시오.

*tnsnames.ora* 파일을 사용하여 TNS를 구현한 경우에는 서버 시스템의 *tnsnames.ora* 파일에서 이 호스트 이름, 포트 및 SID를 얻을 수 있습니다. 자세한 정보를 얻으려면 Oracle 관리자에게 문의하십시오.

5. 테스트 단추를 클릭하여 연결을 테스트하십시오.

### IBM SPSS Modeler에서 Oracle Data Mining Integration 사용

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

도구 > 옵션 > 헬퍼 애플리케이션

2. **Oracle** 탭을 클릭하십시오.

**Oracle Data Mining Integration 사용.** IBM SPSS Modeler 창의 맨 아래에서 데이터베이스 모델링 팔레트를 사용으로 설정(아직 표시되지 않은 경우)하고 Oracle Data Mining 알고리즘에 대한 노드를 추가합니다.

**Oracle 연결.** 유효한 사용자 이름 및 비밀번호와 함께 모델 작성 및 저장에 사용되는 기본 Oracle ODBC 데이터 소스를 지정하십시오. 이 설정은 개별 모델링 노드 및 모델 너겟에서 대체될 수 있습니다.

참고: 모델링을 위해 사용되는 데이터베이스 연결은 데이터에 액세스하는 데 사용되는 연결과 동일하거나 동일하지 않을 수 있습니다. 예를 들어, Oracle 데이터베이스에서 데이터에 액세스하는 스트림이 있는 경우, 정리 또는 기타 조작을 위해 데이터를 IBM SPSS Modeler에 다운로드한 다음 모델링 목적으로 데이터를 다른 Oracle 데이터베이스에 업로드할 수 있습니다. 또는 원래 데이터가 플랫 파일 또는 다른(비Oracle) 소스에 상주할 수 있으

며 이 경우에는 모델링을 위해 데이터를 Oracle에 업로드해야 합니다. 모든 경우에 데이터는 모델링에 사용되는 데이터베이스에서 작성된 임시 테이블에 자동으로 업로드됩니다.

**Oracle Data Mining 모델을 겹쳐쓰려고 할 때 경고.** 데이터베이스에 저장된 모델이 경고 없이 IBM SPSS Modeler에 의해 겹쳐써지지 않게 하려면 이 옵션을 선택하십시오.

**Oracle Data Mining 모델 나열.** 사용 가능한 데이터 마이닝 모델을 표시합니다.

**Oracle Data Miner의 시작 사용(선택사항).** 사용으로 설정되면 IBM SPSS Modeler가 Oracle Data Miner 애플리케이션을 시작하도록 허용합니다. 자세한 정보는 38 페이지의 『Oracle Data Miner』의 내용을 참조하십시오.

**Oracle Data Miner 실행 파일의 경로(선택사항).** Windows용 Oracle Data Miner 실행 파일의 실제 위치를 지정합니다(예: C:\odm\bin\odminerw.exe). Oracle Data Miner는 IBM SPSS Modeler와 함께 설치되지 않으므로 Oracle 웹 사이트(<http://www.oracle.com/technology/products/bi/odm/odminer.html>)에서 올바른 버전을 다운로드하여 클라이언트에서 설치해야 합니다.

SQL 생성 및 최적화 사용

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

도구 > 스트림 특성 > 옵션

2. 탐색 분할창에서 **최적화** 옵션을 클릭하십시오.

3. **SQL 생성** 옵션이 사용으로 설정되어 있는지 확인하십시오. 이 설정은 데이터베이스 모델링이 작동하기 위해 필수입니다.

4. **SQL 생성 최적화 및 기타 실행 최적화**를 선택하십시오(엄격하게 요구되지 않지만 최적화된 성능을 위해서는 강력하게 권장됨).

## Oracle Data Mining을 사용하여 모델 작성

Oracle 모델 작성 노드는 IBM SPSS Modeler의 기타 모델링 노드와 마찬가지로 작동하며 몇 가지 예외가 있습니다. IBM SPSS Modeler 창의 맨 아래에 있는 데이터베이스 모델링 팔레트에서 이 노드에 액세스할 수 있습니다.

데이터 고려사항

Oracle을 사용하려면 범주형 데이터를 문자열 형식(CHAR 또는 VARCHAR2)으로 저장해야 합니다. 결과적으로 IBM SPSS Modeler에서는 측정 수준이 플래그 또는 명목(범주형)인 숫자 저장 영역 필드를 ODM 모델에 대한 입력으로 지정할 수 없습니다. 필요한 경우에는 재분류 노드를 사용하여 IBM SPSS Modeler에서 숫자를 문자열로 변환할 수 있습니다.

**대상 필드.** 하나의 필드만 ODM 분류 모델의 출력(목표) 필드로 선택할 수 있습니다.

**모델 이름.** Oracle 11gR1부터 unique라는 이름은 키워드이므로 사용자 정의 모델 이름으로 사용할 수 없습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

일반 주석

- PMML 내보내기/가져오기는 Oracle Data Mining에 의해 작성된 모델에 대해 IBM SPSS Modeler에서 제공되지 않습니다.
- 모델 스코어링은 항상 ODM 내에서 발생합니다. 데이터가 IBM SPSS Modeler 내에서 시작되거나 준비되어야 하는 경우에는 데이터 세트를 임시 테이블에 업로드해야 할 수 있습니다.
- IBM SPSS Modeler에서는 일반적으로 하나의 예측 및 연관된 확률 또는 신뢰도가 전달됩니다.
- IBM SPSS Modeler는 모델 작성 및 스코어링에서 사용할 수 있는 필드 수를 1,000으로 제한합니다.
- IBM SPSS Modeler는 IBM SPSS Modeler Solution Publisher를 사용하여 실행을 위해 게시된 스트림 내에서 ODM 모델을 스코어링할 수 있습니다.

## Oracle 모형 서버 옵션

모델링에 대한 데이터를 업로드하는 데 사용되는 Oracle 연결을 지정하십시오. 필요한 경우, 각 모델링 노드에 대한 서버 탭에서 연결을 선택하여 헬퍼 애플리케이션 대화 상자에서 지정된 기본 Oracle 연결을 대체할 수 있습니다. 자세한 정보는 [24 페이지의 『Oracle과의 통합 사용』](#)의 내용을 참조하십시오.

### 설명

- 모델링에 사용되는 연결은 스트림에 대한 소스 노드에 사용되는 연결과 동일하거나 동일하지 않을 수 있습니다. 예를 들어, Oracle 데이터베이스에서 데이터에 액세스하는 스트림이 있는 경우, 정리 또는 기타 조작을 위해 데이터를 IBM SPSS Modeler에 다운로드한 다음 모델링 목적으로 데이터를 다른 Oracle 데이터베이스에 업로드할 수 있습니다.
- ODBC 데이터 소스 이름이 각 IBM SPSS Modeler 스트림에 효과적으로 임베드됩니다. 한 호스트에서 작성된 스트림이 다른 호스트에서 실행되는 경우 데이터 소스의 이름이 각 호스트에서 동일해야 합니다. 그렇지 않으면 각 소스 또는 모델링 노드의 서버 탭에서 다른 데이터 소스가 선택될 수 있습니다.

## 오분류 비용

일부 컨텍스트에서는 특정 오차 유형이 다른 유형에 비해 더 값비쌉니다. 예를 들어, 고위험 신용 거래 신청자를 저위험(오차의 한 유형)으로 분류하는 것이 저위험 신청자를 고위험(다른 유형의 오차)으로 분류하는 것보다 비용이 더 나갈 수 있습니다. 오분류 비용을 통해 여러 다른 유형의 예측 오차의 상대적 중요도를 지정할 수 있습니다.

오분류 비용은 기본적으로 특정 결과에 적용된 가중값입니다. 이 가중값은 모델에 영향을 미치는 요인이 되어 실제로 예측변수를 변경합니다(값비싼 실수에 대비하는 보호의 한 방법으로).

C5.0 모델을 제외하고, 오분류 비용은 모델을 스코어링할 때 적용되지 않으며 자동 분류자 노드, 평가 노드 또는 분석 노드를 사용하여 모델을 순위화하거나 비교할 때 고려되지 않습니다. 비용을 포함한 모델은 포함하지 않은 모델보다 적은 오차를 생성하지 않고 전체 정확도 측면에서 더 높게 순위화되지 않을 수는 있지만, 보다 저렴 오차를 선호하는 편향이 내재되어 있어서 실제적으로 성능이 더 좋을 수 있습니다.

비용 교차표는 예측 범주와 실제 범주의 가능한 각 조합의 비용을 표시합니다. 기본적으로 모든 오분류 비용은 1.0으로 설정됩니다. 사용자 정의 비용 값을 입력하려면 **오분류 비용 사용**을 선택하고 비용 교차표에 사용자 정의 값을 입력하십시오.

오분류 비용을 변경하려면 예측 및 실제 값의 원하는 조합에 해당하는 셀을 선택하고 셀의 기존 내용을 삭제한 후 셀의 원하는 비용을 입력하십시오. 비용은 자동으로 대칭되지 않습니다. 예를 들어, A를 B로 오분류한 비용을 2.0으로 설정할 경우 이 설정을 명시적으로 변경하지 않으면 B를 A로 오분류하는 비용의 기본값은 여전히 1.0입니다.

참고: 의사결정 트리 모형에서만 작성 시 비용을 지정할 수 있습니다.

## Oracle Naive Bayes

Naive Bayes는 분류 문제점을 위한 잘 알려진 알고리즘입니다. 이 모델은 제안된 모든 예측 변수를 서로 독립된 것으로 취급하므로 *naïve*라는 이름이 사용됩니다. Naive Bayes는 속성과 목표 속성의 조합에 대한 조건부 확률을 계산하는 빠르고 확장 가능한 알고리즘입니다. 훈련 데이터로부터 독립적인 확률이 설정됩니다. 이 확률은 각 입력 변수에서 각각의 값 범주가 발생하는 경우 각 목표 클래스의 우도를 제공합니다.

- 교차 검증은 모델을 작성하는 데 사용된 동일한 데이터에 대한 모형 정확도를 검정하는 데 사용됩니다. 이는 모델을 작성하는 데 사용할 수 있는 케이스 수가 적은 경우에 특히 유용합니다.
- 모델 출력은 교차표 형식으로 찾을 수 있습니다. 교차표 내의 수는 예측 클래스(열)과 예측변수 값 조합(행)을 연관시키는 조건부 확률입니다.

## Naive Bayes 모형 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**파티션된 데이터 사용.** 파티션 필드가 정의된 경우 이 옵션을 사용하면 학습 파티션의 데이터만 사용하여 모델을 작성합니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

## Naive Bayes 고급 옵션

모델이 작성될 때 학습 데이터에서 지정된 값 또는 쌍이 충분히 발생하지 않으면 개별 예측변수 속성 값 또는 값 쌍은 무시됩니다. 값을 무시하는 임계값은 학습 데이터의 레코드 수를 기준으로 하여 분수로 지정됩니다. 이러한 임계값을 조정하면 잡음이 줄어들고 모델이 기타 데이터 세트로 일반화될 수 있는 기능이 개선됩니다.

- 싱글톤 임계값.** 지정된 예측변수 속성 값에 대한 임계값을 지정합니다. 지정된 값의 발생 수는 지정된 분수 이상이어야 합니다. 그렇지 않으면 값이 무시됩니다.
- 대응별 임계값.** 지정된 속성 및 예측변수 값 쌍에 대한 임계값을 지정합니다. 지정된 값 쌍의 발생 수는 지정된 분수 이상이어야 합니다. 그렇지 않으면 쌍이 무시됩니다.

**예측 확률.** 모델이 대상 필드의 가능한 결과에 대한 올바른 예측의 확률을 포함할 수 있게 합니다. 이 기능을 사용 하려면 선택을 선택하고 지정 단추를 클릭하고 가능한 결과 중 하나를 선택한 후 삽입을 클릭하십시오.

**예측 세트 사용.** 대상 필드의 가능한 모든 결과에 대한 가능한 모든 결과의 테이블을 생성합니다.

## Oracle 적응형 베이스

적응형 Bayes 네트워크(ABN)는 최소 설명 길이(MDL) 및 자동 필드선택을 사용하여 베이지안 네트워크 분류자를 구성합니다. ABN은 Naive Bayes가 적합하지 않은 몇 가지 상황에 더 적합하며 대부분의 기타 상황에서도 성능은 느려질 수 있으나 어느 정도의 성과를냅니다. ABN 알고리즘은 단순화된 의사결정 트리(단일-필드), 가지치기를 한 Naive Bayes 및 증폭된 다기능 모형을 포함하는 세 가지 유형의 고급 베이지안 기반 모델을 작성하기 위한 기능을 제공합니다.

**참고:** Oracle 적합 Bayes 알고리즘은 Oracle 12C에서 삭제되었으며 Oracle 12C를 사용할 때 IBM SPSS Modeler에서 지원되지 않습니다. [http://docs.oracle.com/database/121/DMPRG/release\\_changes.htm#DMPRG726](http://docs.oracle.com/database/121/DMPRG/release_changes.htm#DMPRG726)의 내용을 참조하십시오.

### 생성된 모델

단일 필드 작성 모드에서 ABN은 사람이 읽을 수 있는 규칙 세트를 기반으로 하여 단순화된 의사결정 트리를 작성하며 이를 사용하여 비즈니스 사용자 또는 분석가가 모델의 예측값 및 실제값의 기초를 이해하고 이에 따라 행동하거나 다른 사용자에게 설명할 수 있습니다. 이는 Naive Bayes 및 다기능 모형에 대한 유의적인 장점입니다. 이러한 규칙은 IBM SPSS Modeler에서 표준 규칙 세트처럼 찾을 수 있습니다. 규칙의 단순 세트는 다음과 같이 표시됩니다.

```
IF MARITAL_STATUS = "Married"
AND EDUCATION_NUM = "13-16"
THEN CHURN= "TRUE"
Confidence = .78, Support = 570 cases
```

가지치기된 Naive Bayes 및 다기능 모형은 IBM SPSS Modeler에서 찾을 수 없습니다.

## 적응형 Bayes 모델 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**파티션된 데이터 사용.** 파티션 필드가 정의된 경우 이 옵션을 사용하면 학습 파티션의 데이터만 사용하여 모델을 작성합니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

#### 모델 유형

세 가지 다른 모드의 모델 작성을 선택할 수 있습니다.

- 다기능.** NB 모델, 단일 및 다기능 곱 확률 모델을 포함하여 수많은 모델을 작성하고 비교합니다. 이 모드는 가장 철저한 모드이며 그에 따라 일반적으로 계산에 가장 긴 시간이 걸립니다. 규칙은 단일 필드 모델이 가장 적합한 것으로 판명되는 경우에만 생성됩니다. 다기능 또는 NB 모델이 선택되면, 어떠한 규칙도 생성되지 않습니다.
- 단일 필드.** 규칙 세트를 기준으로 하여 단순한 의사결정 트리를 작성합니다. 각 규칙에는 각 결과와 연관된 확률과 함께 조건이 포함됩니다. 규칙은 상호 배타적이며 사람이 읽을 수 있는 형식으로 제공됩니다. 이는 Naive Bayes 및 다기능 모형에 비해 큰 장점입니다.
- Naive Bayes.** 단일 NB 모델을 작성하고 글로벌 표본 사전확률(글로벌 표본에서 목표 값의 분포)과 비교합니다. NB 모델은 글로벌 사전확률보다 더 나은 목표 값의 예측변수인 것으로 판명된 경우에만 출력으로 생성됩니다. 그렇지 않으면 어떠한 모델도 출력으로 생성되지 않습니다.

## 적응형 Bayes 고급 옵션

실행 시간 제한. 분 단위로 최대 작성 시간을 지정하려면 이 옵션을 선택하십시오. 그러면 결과 모델이 덜 정확하더라도 더 짧은 시간에 모델을 작성할 수 있습니다. 모델링 프로세스의 각 마일스톤에서 계속 진행하기 전에 알고리즘이 지정된 시간 동안 다음 마일스톤을 완료할 수 있는지 여부를 검사하여 한계에 도달하면 가장 적합한 모델을 리턴합니다.

**최대 예측자.** 이 옵션을 사용하면 모델의 복잡도를 제한하고 사용되는 예측자 수를 제한하여 성능을 개선할 수 있습니다. 예측자는 모델에 포함된 우도 측도로서 목표에 대한 상관관계의 MDL 측도를 기반으로 하여 순위가 결정됩니다.

**최대 Naive Bayes 예측자.** 이 옵션은 Naive Bayes 모델에서 사용할 예측자의 최대수를 지정합니다.

## Oracle 지원 벡터 머신(SVM)

지원 벡터 머신(SVM)은 데이터 과적합 없이도 예측 정확도를 최대화하기 위해 머신 학습 이론을 사용하는 분류 및 회귀분석 알고리즘입니다. SVM은 학습 데이터의 선택적 비선형 변환을 사용하며 변환된 데이터에서 회귀분석 방정식을 검색하여 범주형 대상의 경우 클래스를 분할하고 연속형 대상의 경우 목표를 적합하게 만듭니다. Oracle의 SVM 구현을 사용하면 두 가지 사용 가능한 커널, 즉, 선형 또는 Gaussian 중 하나를 사용하여 모델을 작성할 수 있습니다. 선형 커널은 비선형 변환을 모두 생략하므로 결과 모형은 본질적으로 회귀 모형입니다.

자세한 정보는 Oracle 데이터 마이닝 애플리케이션 개발자 안내서 및 Oracle 데이터 마이닝 개념을 참조하십시오.

## Oracle SVM 모형 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 Oracle Data Mining 개념을 참조하십시오.

**활성 학습.** 큰 작성 세트를 다루는 방법을 제공합니다. 이 알고리즘은 활성 학습을 함께 사용하여 작은 표본을 기반으로 하는 초기 모델을 작성한 다음 이를 완전한 학습 데이터 세트에 적용하고 결과를 기반으로 하여 표본 및 모델을 점증적으로 적용하십시오. 학습 데이터에 대한 모델이 수렴되거나 허용되는 지원 벡터의 최대 수에 도달 할 때까지 순환이 반복됩니다.

**커널 함수. 선형** 또는 **Gaussian**을 선택하거나 시스템이 가장 적합한 커널을 선택할 수 있도록 기본값인 **시스템 정의**를 사용하십시오. Gaussian kernel은 더 복잡한 관계를 학습할 수 있으나 일반적으로 계산에 시간이 더 오래 걸립니다. 선형 커널로 시작하여 선형 커널이 적합하지 않은 경우에만 Gaussian kernel을 시도해 볼 수 있습니다. 커널 선택이 더 중요한 회귀 모형에서 이런 경우가 더 빈번합니다. 또한 Gaussian kernel을 사용하여 작성된 SVM 모형은 IBM SPSS Modeler에서 찾아 볼 수 없습니다. 선형 커널을 사용하여 작성된 모형은 표준 회귀 모형과 동일한 방법으로 IBM SPSS Modeler에서 찾아볼 수 있습니다.

**정규화 방법.** 연속형 입력 및 대상 필드에 대한 정규화 방법을 지정합니다. 표준화 점수, 최소-최대 또는 지정않음 을 선택할 수 있습니다. 자동 데이터 준비 선택란이 선택된 경우 Oracle은 자동으로 정규화를 수행합니다. 수동으로 정규화 방법을 선택하려면 이 선택란을 선택 취소하십시오.

## Oracle SVM 고급 옵션

**커널 캐시 크기.** 작성 작업 동안 계산된 커널을 저장하는 데 사용할 캐시의 크기를 바이트 단위로 지정합니다. 예상대로 일반적으로 캐시가 크면 빨리 작성됩니다. 기본값은 50MB입니다.

**수렴 허용.** 모델 작성 종료 전에 허용되는 공차 값을 지정합니다. 값은 0 - 1 사이여야 하며 기본값은 0.001입니다. 값이 크면 더 빨리 작성되는 경향이 있으나 덜 정확한 모델이 작성될 수 있습니다.

**표준 편차 지정.** Gaussian kernel에 의해 사용되는 표준 편차 모수를 지정합니다. 이 모수는 모델 복잡도 사이의 균형 및 기타 데이터 세트로 일반화하는 기능(데이터 과적합 및 과소적합)에 영향을 미칩니다. 표준 편차 값이 높으면 과소적합되는 경향이 있습니다. 기본적으로 이 모수는 학습 데이터에서 추정됩니다.

**엡실론 지정.** 회귀 모형에 대해서만 엡실론 집중 모형을 작성할 때 허용되는 오류 구간 값을 지정합니다. 즉, 큰 오류(무시되지 않음)에서 작은 오류(무시됨)를 구분합니다. 값은 0 - 1 사이여야 하며 기본적으로 학습 데이터에서 추정됩니다.

**복잡도 요인 지정.** 학습 데이터에서 측정된 모델 오류 및 데이터 과적합 또는 과소적합을 방지하기 위한 모델 복잡도의 균형을 유지하는 복잡도 요인을 지정합니다. 높은 값은 데이터 과적합의 위험도를 높이면서 오류에 높은 패널티를 부여하며 낮은 값은 오류에 낮은 패널티를 부여하며 과소적합 발생 가능성이 있습니다.

**이상치 비율 지정.** 학습 데이터에서 원하는 이상값 비율을 지정합니다. One-Class SVM 모델에 대해서만 유효합니다. **복잡도 요인 지정** 설정과 함께 사용할 수 없습니다.

**예측 확률.** 모델이 대상 필드의 가능한 결과에 대한 올바른 예측의 확률을 포함할 수 있게 합니다. 이 기능을 사용하려면 선택을 선택하고 **지정** 단추를 클릭하고 가능한 결과 중 하나를 선택한 후 **삽입**을 클릭하십시오.

**예측 세트 사용.** 대상 필드의 가능한 모든 결과에 대한 가능한 모든 결과의 테이블을 생성합니다.

## Oracle SVM 가중치 옵션

분류 모델에서 가중치를 사용하면 가능한 다양한 목표 값의 상대적 중요도를 지정할 수 있습니다. 학습 데이터의 데이터 점이 범주 사이에 현실적으로 분배되지 않은 경우에 유용합니다. 가중치를 사용하면 데이터에서 제대로 표시되지 않는 범주에 대해 보완할 수 있도록 모델을 편향시킬 수 있습니다. 목표 값에 대한 가중치를 늘리는 경우에는 해당 범주에 대한 올바른 예측의 백분율을 늘려야 합니다.

세 가지 방법으로 가중치를 설정할 수 있습니다.

- **학습 데이터 기준.** 기본값입니다. 가중치는 학습 데이터에 있는 범주의 상대적 빈도를 기반으로 합니다.
- **모든 클래스에 대해 동등함.** 모든 범주에 대한 가중치는  $1/k$ 로 정의됩니다. 여기서  $k$ 는 목표 범주의 수입니다.
- **사용자 정의.** 자체 가중치를 지정할 수 있습니다. 가중치의 시작 값은 모든 클래스에 대해 동일하게 설정됩니다. 개별 범주에 대한 가중치를 사용자 정의 값으로 조정할 수 있습니다. 특정 범주의 가중치를 조정하려면 원하는 범주에 해당하는 테이블의 가중치 셀을 선택한 후 셀의 내용을 삭제하고 원하는 값을 입력하십시오.

모든 범주에 대한 가중치의 합계는 1.0이어야 합니다. 합계가 1.0이 아니면 값을 자동으로 표준화하는 옵션과 함께 경고 메시지가 표시됩니다. 이 자동 조정은 가중치 제한 조건을 적용하면서 범주에서 비율을 유지합니다. 언제든지 **표준화** 단추를 클릭해서 이 조정을 수행할 수 있습니다. 모든 범주에 동일한 값으로 테이블을 재설정하려면 **평준화** 단추를 클릭하십시오.

## Oracle 일반화 선형 모형(GLM)

(11g 전용) 일반화 선형 모형은 선형 모형에 의해 작성된 제한적인 가정을 완화시킵니다. 여기에는 목표변수가 정규 분포를 가지며 목표변수에 대한 예측자의 영향이 본질적으로 선형이라는 가정 등이 포함됩니다. 일반화 선형 모형은 다항분포 또는 포아송 분포와 같이 목표의 분포가 비명목 분포를 갖기 쉬운 예측에 적합합니다. 이와 유사하게 일반화 선형 모형은 예측자 및 목표 사이의 관계 또는 링크가 비선형이 되기 쉬운 경우에 유용합니다.

자세한 정보는 *Oracle* 데이터 마이닝 애플리케이션 개발자 안내서 및 *Oracle* 데이터 마이닝 개념을 참조하십시오.

### Oracle GLM 모형 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: *CustomerID*)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(*Oracle Adaptive Bayes*, *Oracle O-Cluster* 및 *Oracle Apriori* 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

**정규화 방법.** 연속형 입력 및 대상 필드에 대한 정규화 방법을 지정합니다. 표준화 점수, 최소-최대 또는 지정않음 을 선택할 수 있습니다. 자동 데이터 준비 선택란이 선택된 경우 Oracle은 자동으로 정규화를 수행합니다. 수동으로 정규화 방법을 선택하려면 이 선택란을 선택 취소하십시오.

**결측값 처리.** 입력 데이터에서 결측값을 처리하는 방법을 지정합니다.

- 평균 또는 최빈값으로 바꾸기는 수치 속성의 결측값을 평균 값으로 대체하고 범주형 속성의 결측값을 최빈값으로 대체합니다.
- 완전한 레코드만 사용은 결측값이 있는 레코드를 무시합니다.

### Oracle GLM 고급 옵션

**행 가중치 사용.** 행에 대한 가중치 요인을 포함하는 열을 선택할 수 있는 인접 드롭 다운 목록을 활성화하려면 이 선택란을 선택하십시오.

**행 진단을 테이블에 저장.** 행 수준의 진단을 포함하는 테이블의 이름을 입력할 수 있는 인접 텍스트 필드를 활성화하려면 이 선택란을 선택하십시오.

**계수 신뢰수준.** 목표에 대해 예측되는 값이 모델에 의해 계산된 신뢰구간 내에 있게 되는 0.0에서 1.0까지의 정확도입니다. 신뢰한계는 계수 통계량과 함께 리턴됩니다.

**목표에 대한 참조범주.** 사용자 정의를 선택하여 참조범주로 사용할 대상 필드에 대한 값을 선택하거나 기본값인 자동으로 두십시오.

**능선 회귀.** 능선 회귀는 변수의 상관관계 차수가 너무 높은 상황을 보완하는 기술입니다. 자동 옵션을 사용하여 알고리즘이 이 기술 사용을 제어하도록 하거나 사용할 수 없음 및 사용 옵션을 사용하여 수동으로 제어할 수 있습니다. 수동으로 능선 회귀를 사용하도록 선택한 경우, 인접 필드에 값을 입력하여 능선 모수에 대한 시스템 기본 값을 대체할 수 있습니다.

**능선 회귀에 대한 VIF 생성.** 선형 회귀에 능선이 사용되는 경우 분산 팽창 계수(VIF) 통계를 생성하려면 이 상자를 선택하십시오.

**예측 확률.** 모델이 대상 필드의 가능한 결과에 대한 올바른 예측의 확률을 포함할 수 있게 합니다. 이 기능을 사용하려면 선택을 선택하고 지정 단추를 클릭하고 가능한 결과 중 하나를 선택한 후 삽입을 클릭하십시오.

**예측 세트 사용.** 대상 필드의 가능한 모든 결과에 대한 가능한 모든 결과의 테이블을 생성합니다.

## Oracle GLM 가중치 옵션

분류 모델에서 가중치를 사용하면 가능한 다양한 목표 값의 상대적 중요도를 지정할 수 있습니다. 학습 데이터의 데이터 점이 범주 사이에 현실적으로 분배되지 않은 경우에 유용합니다. 가중치를 사용하면 데이터에서 제대로 표시되지 않는 범주에 대해 보완할 수 있도록 모델을 편향시킬 수 있습니다. 목표 값에 대한 가중치를 늘리는 경우에는 해당 범주에 대한 올바른 예측의 백분율을 늘려야 합니다.

세 가지 방법으로 가중치를 설정할 수 있습니다.

- **학습 데이터 기준.** 기본값입니다. 가중치는 학습 데이터에 있는 범주의 상대적 빈도를 기반으로 합니다.
- **모든 클래스에 대해 동등함.** 모든 범주에 대한 가중치는  $1/k$ 로 정의됩니다. 여기서  $k$ 는 목표 범주의 수입니다.
- **사용자 정의.** 자체 가중치를 지정할 수 있습니다. 가중치의 시작 값은 모든 클래스에 대해 동일하게 설정됩니다. 개별 범주에 대한 가중치를 사용자 정의 값으로 조정할 수 있습니다. 특정 범주의 가중치를 조정하려면 원하는 범주에 해당하는 테이블의 가중치 셀을 선택한 후 셀의 내용을 삭제하고 원하는 값을 입력하십시오.

모든 범주에 대한 가중치의 합계는 1.0이어야 합니다. 합계가 1.0이 아니면 값을 자동으로 표준화하는 옵션과 함께 경고 메시지가 표시됩니다. 이 자동 조정은 가중치 제한조건을 적용하면서 범주에서 비율을 유지합니다. 언제든지 표준화 단추를 클릭해서 이 조정을 수행할 수 있습니다. 모든 범주에 동일한 값으로 테이블을 재설정하려면 평준화 단추를 클릭하십시오.

## Oracle 의사결정 트리

Oracle 데이터 마이닝은 일반적인 분류 및 회귀분석 트리 알고리즘을 기반으로 하여 클래식 의사결정 트리 기능을 제공합니다. ODM 의사결정 트리 모형에는 신뢰도, 지원 및 분할 기준을 포함하여 각 노드에 대한 완전한 정보가 포함됩니다. 각 노드에 대한 전체 규칙이 표시될 수 있으며 결측값이 있는 케이스에 모델을 적용할 때 대체하여 사용할 수 있도록 각 노드에 대한 대용 속성이 제공됩니다.

의사결정 트리는 광범위하게 적용될 수 있으며 적용 및 이해가 쉽기 때문에 널리 사용됩니다. 의사결정 트리는 가장 적합한 "분할자", 즉, 다운스트림 데이터 레코드를 더 동일한 모집단으로 분할하는 속성 절단점(AGE > 55 등)을 검색하여 각각의 잠재적인 입력 속성을 조사합니다. 각 분할 의사결정 후에는 ODM이 전체 트리까지 확장하면서 유사한 레코드, 항목 또는 사람 모집단을 나타내는 터미널 "리프"를 작성하면서 프로세스를 반복합니다. 루트 트리 노드(총계 모집단 등)에서 아래로 검색하면서 의사결정 트리가 사람이 읽을 수 있는 IF A, then B 문 규칙을 제공합니다. 이러한 의사결정 트리 규칙은 각 트리 노드에 대한 지원 및 신뢰도도 제공합니다.

적응형 Bayes 네트워크는 각 예측에 대한 설명을 제공하는데 유용한 짧고 단순한 규칙을 제공할 수 있는 반면에 의사결정 트리는 각 분할 의사결정에 대한 전체 Oracle 데이터 마이닝 규칙을 제공합니다. 의사결정 트리는 최고의 고객, 건강한 환자, 사기와 연관된 요소 등에 대한 세부사항 프로파일을 개발하는 데도 유용합니다.

## 의사결정 트리 모형 옵션

모델 이름, 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 Oracle Data Mining 개념을 참조하십시오.

**불순도 메트릭.** 각 노드에서 데이터를 분할하기 위한 최고의 검정 질문을 찾기 위해 사용할 메트릭을 지정합니다. 최고의 분할자 및 분할 값은 노드 내의 엔티티에 대해 목표 값 동질성에서 가장 큰 증가를 발생시키는 분할자 및 분할 값입니다. 동질성은 메트릭에 따라 측정됩니다. 지원되는 메트릭은 지니 및 엔트로피입니다.

## 의사결정 트리 고급 옵션

**최대 깊이.** 작성할 트리 모델의 최대 깊이를 설정합니다.

**노드 내의 레코드의 최소 퍼센트.** 노드당 최소 레코드 수의 퍼센트를 설정합니다.

**분할에 대한 레코드의 최소 퍼센트.** 모델을 학습시키기 위해 사용되는 레코드의 총 수의 퍼센트로 표시되는 상위 노드 내의 최소 레코드 수를 설정합니다. 레코드 수가 이 퍼센트 미만이면 분할이 시도되지 않습니다.

**노드 내의 최소 레코드.** 리턴되는 최소 레코드 수를 설정합니다.

**분할에 대한 최소 레코드 수.** 값으로 표시되는 상위 노드 내의 최소 레코드 수를 설정합니다. 레코드 수가 이 값 미만이면 분할이 시도되지 않습니다.

**규칙 식별자.** 선택하면 특정 분할이 작성되는 트리 내의 노드를 식별하기 위한 문자열이 모델에 포함됩니다.

**예측 확률.** 모델이 대상 필드의 가능한 결과에 대한 올바른 예측의 확률을 포함할 수 있게 합니다. 이 기능을 사용하려면 선택을 선택하고 **지정** 단추를 클릭하고 가능한 결과 중 하나를 선택한 후 **삽입**을 클릭하십시오.

**예측 세트 사용.** 대상 필드의 가능한 모든 결과에 대한 가능한 모든 결과의 테이블을 생성합니다.

## Oracle O-Cluster

Oracle O-군집 알고리즘은 데이터 모집단 내에서 자연적으로 발생하는 집단을 식별합니다. 직교 파티셔닝 군집(O-군집)은 계층 구조 눈금 기반의 군집 모델을 작성하는 Oracle 독점 군집 알고리즘입니다. 즉, 입력 속성 공간에 축 병렬(직교) 파티션을 작성합니다. 알고리즘은 회귀적으로 작동합니다. 결과 계층 구조는 군집에 속성 공간을 바둑판 모양으로 배열하는 불규칙적인 눈금을 표시합니다.

O-군집 알고리즘은 수치 및 범주형 속성을 둘 다 처리하며 ODM은 자동으로 가장 적합한 군집 정의를 선택합니다. ODM은 군집 세부사항 정보, 군집 규칙, 군집 중심값을 제공하며 소속군집에 대한 모집단을 스코어링하는 데 사용될 수 있습니다.

### 0-군집 모형 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

**최대 군집 수.** 생성되는 군집의 최대 수를 설정합니다.

### 0-군집 고급 옵션

**최대 버퍼.** 최대 버퍼 크기를 설정합니다.

**민감도.** 새 군집을 분할하는 데 필요한 최대 밀도를 지정하는 분수를 설정합니다. 이 분수는 글로벌 균일 밀도와 연관됩니다.

## Oracle K-평균

Oracle K-평균 알고리즘은 데이터 모집단 내에서 자연적으로 발생하는 집단을 식별합니다. K-평균 알고리즘은 데이터를 미리 결정된 군집 수로 분할하는 거리 기반의 군집 알고리즘입니다. 단, 충분한 구분 케이스가 있어야 합니다. 거리 기반의 알고리즘은 데이터 점 사이의 유사성을 측정하기 위해 거리 메트릭(함수)에 의존합니다. 데이터 점은 사용된 거리 메트릭에 따라 가장 가까운 군집에 지정됩니다. ODM은 K-평균의 개선된 버전을 제공합니다.

K-평균 알고리즘은 계층적 군집을 지원하며 수치 및 범주형 속성을 처리하며 모집단을 사용자가 지정한 군집 수로 분할합니다. ODM은 군집 세부사항 정보, 군집 규칙, 군집 중심값을 제공하며 소속군집에 대한 모집단을 스코어링하는 데 사용될 수 있습니다.

## K-평균 모형 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

고유 필드. 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

참고: 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

자동 데이터 준비. (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

군집 수. 생성되는 군집의 수를 설정합니다.

거리 함수. K-평균 군집에 사용할 거리 함수를 지정합니다.

분할 기준. K-평균 군집에 사용할 분할 기준을 지정합니다.

정규화 방법. 연속형 입력 및 대상 필드에 대한 정규화 방법을 지정합니다. 표준화 점수, 최소-최대 또는 지정않음 을 선택할 수 있습니다.

## K-평균 고급 옵션

반복. K-평균 알고리즘에 대한 반복계산 수를 설정합니다.

수렴허용. K-평균 알고리즘에 대한 수렴허용을 설정합니다.

구간 수. K-평균에 의해 설정된 속성 히스토그램 내의 구간 수를 지정합니다. 각 속성에 대한 구간 경계는 전체 학습 데이터 세트에 대해 글로벌 형식으로 계산됩니다. 구간화 방법은 동등 간격입니다. 모든 속성이 동일한 구간 수를 가지나 구간이 하나뿐인 단일값이 있는 속성은 예외입니다.

블록 성장. 군집 데이터를 보유하기 위해 할당된 메모리에 대한 성장 요인을 설정합니다.

최소 퍼센트 속성 지원. 속성이 군집에 대한 규칙 설명에 포함되기 위해 널이 아니어야 하는 속성 값의 분수를 설정합니다. 결측값이 있는 데이터에서 모수값을 너무 높게 설정하면 매우 짧거나 짧거나 비어 있는 규칙이 생성될 수 있습니다.

## Oracle 비음수 교차표 분해(NMF)

비음수 교차표 분해(NMF)는 큰 데이터 세트를 대표적인 속성으로 축소하는 데 유용합니다. NMF는 비선형 주성 분분석(PCA)과 개념은 유사하나 많은 수의 속성을 처리할 수 있으며 다양한 유스 케이스에 대해 사용할 수 있는 강력한 최신 기술의 데이터 마이닝 알고리즘입니다.

NMF는 많은 양의 데이터를 축소하는 데 사용할 수 있습니다. 예를 들어, 텍스트 데이터를 데이터의 차원을 축소하는 더 작고 더 희박한 표시로 축소할 수 있습니다. 즉, 훨씬 적은 수의 변수를 사용하여 동일한 정보를 보유할 수 있습니다. NMF 모델의 출력은 SVM과 같은 감독되는 학습 기술 또는 군집 기술과 같이 감독되지 않는 학습 기술을 사용하여 분석할 수 있습니다. Oracle 데이터 마이닝은 NMF 및 SVM 알고리즘을 사용하여 비정형 텍스트 데이터를 마이닝합니다.

## NMF 모형 옵션

모델 이름. 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

고유 필드. 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

참고: 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

**정규화 방법.** 연속형 입력 및 대상 필드에 대한 정규화 방법을 지정합니다. 표준화 점수, 최소-최대 또는 지정않음 을 선택할 수 있습니다. 자동 데이터 준비 선택란이 선택된 경우 Oracle은 자동으로 정규화를 수행합니다. 수동으로 정규화 방법을 선택하려면 이 선택란을 선택 취소하십시오.

## NMF 고급 옵션

**변수의 수 지정.** 내보낼 변수 수를 지정합니다.

**난수 시드.** NMF 알고리즘에 대한 난수 시드를 설정합니다.

**반복계산 수.** NMF 알고리즘에 대한 반복계산 수를 설정합니다.

**수렴허용.** NMF 알고리즘에 대한 수렴허용을 설정합니다.

**모든 기능 표시.** 최고 기능에 대한 값만 표시되지 않고 모든 기능에 대한 기능 ID 및 신뢰도가 표시됩니다.

## Oracle Apriori

Apriori 알고리즘은 데이터의 연관 규칙을 검색합니다. 예를 들어, "면도기와 애프터쉐이브 로션을 구매하는 고객은 다음 번 구매 시에 면도용 크림을 구매할 수 있습니다." 연관 마이닝 문제점은 두 개의 하위 문제점으로 분해할 수 있습니다.

- 지원이 최소 지원보다 큰 항목의 모든 조합(빈번 항목 세트)을 발견합니다.
- 빈번 항목 세트를 사용하여 원하는 규칙을 생성합니다. 예를 들어, ABC 및 BC가 빈번한 경우에 support(ABC) 대 support(BC)의 비율이 최소한 최소 신뢰도만큼 크면 "A가 BC를 암시한다"라는 규칙이 유지됩니다. ABCD가 빈번하므로 규칙이 최소한의 지원을 가집니다. ODM 연관은 단일 후향 규칙(ABC가 D를 암시함)만 지원합니다.

빈번한 항목 세트의 수는 최소 지원 모수에 의해 제어됩니다. 생성되는 규칙의 수는 빈번한 항목 세트의 수 및 신뢰도 모수에 의해 제어됩니다. 신뢰도 모수가 너무 높게 설정되면 연관 모델에 빈번한 항목 세트는 있으나 규칙이 없을 수 있습니다.

ODM은 Apriori 알고리즘의 SQL 기반 구현을 사용합니다. 후보 생성 및 지원 개수 단계가 SQL 쿼리를 사용하여 구현됩니다. 특화된 인메모리 데이터 구조는 사용되지 않습니다. SQL 쿼리는 미세하게 조정되어 다양한 힌트를 사용하여 데이터베이스 서버에서 효과적으로 실행됩니다.

## Apriori 필드 옵션

모든 모델링 노드에는 필드 탭이 있으며, 여기에서 모델 작성 시 사용할 필드를 지정할 수 있습니다.

Apriori 모형을 작성하려면 먼저 연관 모델링에 관련 항목으로 사용할 필드를 지정해야 합니다.

**유형 노드 설정 사용.** 이 옵션은 업스트림 유형 노드의 필드 정보를 사용하도록 노드에 지시합니다. 기본값입니다.

**사용자 정의 설정 사용.** 이 옵션은 업스트림 유형 노드에 제공된 정보 대신 여기에 지정된 필드 정보를 사용하도록 노드에 알립니다. 이 옵션을 선택한 후 트랜잭션 형식을 사용 중인지 여부에 따라 대화 상자에서 나머지 필드를 지정하십시오.

트랜잭션 형식을 사용하지 않는 경우 다음을 지정하십시오.

- **입력.** 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 입력으로 설정하는 것과 유사합니다.
- **파티션.** 이 필드에서는 모델 작성에 대한 학습, 검정, 검증 단계에 대한 별도의 샘플로 데이터를 파티션하는 데 사용되는 필드를 지정할 수 있습니다.

트랜잭션 형식을 사용하는 경우 다음을 지정하십시오.

**트랜잭션 형식 사용.** 데이터를 항목당 행에서 케이스당 행으로 변환하려면 이 옵션을 사용하십시오.

이 옵션을 선택하면 이 대화 상자의 아래 부분에서 필드 제어가 변경됩니다.

트랜잭션 형식의 경우 다음을 지정하십시오.

- **ID.** 목록에서 ID 필드를 선택합니다. 숫자 또는 기호 필드를 ID 필드로 사용할 수 있습니다. 이 필드의 각 고유한 값은 분석의 특정 단위를 표시해야 합니다. 예를 들어, 장바구니 애플리케이션에서 각 ID는 단일 고객을 나타낼 수 있습니다. 웹 로그 분석 애플리케이션의 경우 각 ID는 컴퓨터(IP 주소별) 또는 사용자(로그인 데이터별)를 나타낼 수 있습니다.
- **내용.** 모델에 대한 내용 필드를 지정하십시오. 이 필드는 연관 모델링에서 관심이 있는 항목을 포함합니다.
- **파티션.** 이 필드에서는 모델 작성에 대한 학습, 검정, 검증 단계에 대한 별도의 샘플로 데이터를 파티션하는 데 사용되는 필드를 지정할 수 있습니다. 하나의 샘플을 사용하여 모델을 생성하고, 다른 샘플로 검정하면, 현재 데이터와 유사한 더 큰 데이터 세트에 대해 모델을 일반화할 때 효율성을 효과적으로 표시할 수 있습니다. 유형 또는 파티션 노드를 사용하여 다중 파티션 필드를 정의한 경우 파티셔닝을 사용하는 각 모델링 노드에 있는 필드 탭에서 단일 파티션 필드를 선택해야 합니다. (하나의 파티션만 존재하는 경우 파티션이 사용될 때마다 자동으로 사용됩니다.) 또한 분석에서 선택한 파티션을 적용하려면 노드의 모델 옵션 탭에서 파티셔닝을 사용 가능하게 해야 합니다. (이 옵션을 선택 취소하면 필드 설정을 변경하지 않고도, 파티셔닝을 사용하지 않을 수 있습니다.)

## Apriori 모형 옵션

모델 이름, 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

**최대 규칙 길이.** 규칙의 최대 전제조건 수를 2에서 20까지의 정수로 설정하십시오. 이는 규칙의 복잡도를 제한하기 위한 방법입니다. 규칙이 너무 복잡하거나 너무 세밀하거나 규칙 세트의 학습 시간이 너무 오래 걸리면 이 설정을 줄여 보십시오.

**최소 신뢰도.** 최소 신뢰수준을 0에서 1 사이로 설정하십시오. 신뢰도가 지정된 기준보다 낮은 규칙은 삭제됩니다.

**최소 지원.** 최소 지원 임계값을 0에서 1 사이의 값으로 설정하십시오. Apriori는 빈도가 최소 지원 임계값을 초과하는 패턴을 검색합니다.

## Oracle 최소 설명 길이(MDL)

Oracle 최소 설명 길이(MDL) 알고리즘은 목표 속성에 대해 가장 큰 영향을 가진 속성을 식별하는 데 도움을 줍니다. 가장 큰 영향을 가진 속성을 알면 비즈니스를 더 잘 이해하고 관리하는 데 도움이 되며 모델링 활동을 단순화하는 데 도움이 되는 경우가 많습니다. 또한 이러한 속성은 모델을 강화하기 위해 추가하고자 하는 데이터 유형을 표시할 수 있습니다. 예를 들어, MDL은 제조업체 파트의 품질 예측, 이탈과 연관된 요인 또는 특정 질병의 치치와 가장 밀접하게 연관된 세균 등과 관련된 프로세스 속성을 찾는데 사용될 수 있습니다.

Oracle MDL은 목표를 예측하는 데 중요하지 않은 것으로 간주되는 입력 필드를 삭제합니다. 그런 다음 나머지 입력 필드를 사용하여 Oracle Data Miner에서 볼 수 있는 Oracle 모델과 연관된 세분화되지 않은 모델 너깃을 작성합니다. Oracle Data Miner에서 모델을 찾아보면 나머지 입력 필드를 표시하는 차트가 표시되고 목표 예측에 중요한 순서대로 순위가 매겨집니다.

음수 순위화는 잡음을 표시합니다. 0 또는 그 미만의 값으로 순위가 지정된 입력 필드는 예측에 기여하지 않으며 데이터에서 제거되어야 합니다.

차트를 표시하려면 다음을 수행하십시오.

1. 모델 팔레트에서 세분화되지 않은 모델 너깃을 마우스 오른쪽 단추로 클릭하고 **찾아보기**를 선택하십시오.
2. 모델 창에서 단추를 클릭하여 Oracle Data Miner를 실행하십시오.

3. Oracle Data Miner에 연결하십시오. 자세한 정보는 [38 페이지의 『Oracle Data Miner』](#)의 내용을 참조하십시오.
4. Oracle Data Miner 네비게이터 패널에서 모델을 확장한 다음 속성 중요도를 선택하십시오.
5. 관련된 Oracle 모델을 선택하십시오. IBM SPSS Modeler에서 지정한 대상 필드와 이름이 동일합니다. 올바른 것인지 확신할 수 없으면 속성 중요도 풀더를 선택하고 작성 날짜별로 모델을 검색하십시오.

## MDL 모형 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**고유 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: *CustomerID*)가 될 수 있습니다. IBM SPSS Modeler는 이 키 필드가 숫자여야 한다는 제한사항을 부과합니다.

**참고:** 이 필드는 모든 Oracle 노드에 대해 선택적입니다(Oracle Adaptive Bayes, Oracle O-Cluster 및 Oracle Apriori 제외).

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

## Oracle 속성 중요도(AI)

속성 중요도의 목적은 데이터 세트 내의 어떤 속성이 결과와 연관되는지, 최종 결과에 어느 정도로 영향을 미치는지 알아내는 것입니다. Oracle 속성 중요도 노드는 데이터를 분석하고 패턴을 찾고 연관된 신뢰도 수준의 결과를 예측합니다.

## AI 모델 옵션

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**파티션된 데이터 사용.** 파티션 필드가 정의된 경우 이 옵션을 사용하면 학습 파티션의 데이터만 사용하여 모델을 작성합니다.

**자동 데이터 준비.** (11g 전용) Oracle Data Mining의 자동화된 데이터 준비 모드를 사용(기본값) 또는 사용 안함으로 설정합니다. 이 선택란이 선택되면 ODM이 알고리즘에 필요한 데이터 변환을 자동으로 수행합니다. 자세한 정보는 *Oracle Data Mining* 개념을 참조하십시오.

## AI 선택 옵션

옵션 탭으로 모델 너깃에 입력 필드를 선택하거나 제외시키기 위한 기본 설정을 지정할 수 있습니다. 그런 다음 모델을 스트림에 추가하여 후속 모델 작성 노력에 사용할 필드의 서브세트를 선택할 수 있습니다. 또는 모델을 생성한 후 모델 브라우저에서 추가 필드를 선택 또는 선택 취소하여 이 설정을 대체할 수 있습니다. 하지만 기본 설정은 추가로 변경하지 않고도 모델 너깃을 적용할 수 있어서 특히 스크립팅 용도에 유용할 수 있습니다.

다음 옵션을 사용할 수 있습니다.

**순위 지정된 모든 필드.** 중요, 보통 또는 중요하지 않음과 같은 순위를 기준으로 하여 필드를 선택합니다. 한 순위 또는 또 다른 순위에 레코드를 지정하는 데 사용하는 절사 값 외에 각 순위의 레이블을 편집할 수 있습니다.

**최대 필드 수.** 중요도에 따라 상위 *n*개의 필드를 선택합니다.

**다음보다 큰 중요도.** 중요도가 지정된 값보다 큰 모든 필드를 선택합니다.

대상 필드는 선택과 상관 없이 항상 보존됩니다.

## AI 모델 너깃 모델 탭

Oracle AI 모델 너깃의 모델 탭에서는 모든 입력의 순위 및 중요도를 표시하고, 왼쪽에 있는 열의 선택란을 사용하여 필터링을 위해 필드를 선택할 수 있습니다. 스트림을 실행할 때 목표 예측과 함께 선택된 필드만 유지됩니다.

다. 기타 입력 필드는 삭제됩니다. 기본 선택은 모델링 노드에 지정된 옵션에 기반하지만, 필요에 따라 추가 필드를 선택하거나 선택 취소할 수 있습니다.

- 순위, 필드 이름, 중요도 또는 기타 표시된 열로 목록을 정렬하려면 열 헤더를 클릭하십시오. 또는 정렬기준 단추 옆의 목록에서 원하는 항목을 선택하고 위로 및 아래로 화살표를 사용하여 정렬 방향을 변경하십시오.
- 도구 모음을 사용하여 모든 필드를 선택 또는 선택 취소하고 필드 확인 대화 상자에 액세스할 수 있습니다. 이 대화 상자에서는 순위 또는 중요도를 기준으로 필드를 선택할 수 있습니다. 또한 Shift 또는 Ctrl 키를 누른 상태로 필드를 클릭하여 선택을 확장할 수 있습니다.
- 중요, 주변 또는 중요하지 않음으로 입력을 순위화할 때 임계값은 테이블 아래 범례에 표시됩니다. 이러한 값은 모델링 노드에 지정됩니다.

## Oracle 모델 관리

Oracle 모델은 기타 IBM SPSS Modeler 모델과 동일한 방법으로 모델 팔레트에 추가될 수 있으며 매우 유사한 방법으로 사용될 수 있습니다. 단, IBM SPSS Modeler에서 작성된 각 Oracle 모델이 데이터베이스 서버에 저장된 모델을 실제로 참조하는 경우, 몇 가지 중요한 차이가 있습니다.

### Oracle 모델 너깃 서버 탭

IBM SPSS Modeler를 통해 ODM 모델을 작성하면 IBM SPSS Modeler에서 모델이 작성되고 Oracle 데이터베이스에서 모델이 작성되거나 교체됩니다. 이런 종류의 IBM SPSS Modeler 모델은 데이터베이스 서버에 저장된 데이터베이스 모델의 내용을 참조합니다. IBM SPSS Modeler는 동일하게 작성된 모델 키 문자열을 IBM SPSS Modeler 모델 및 Oracle 모델 둘 다에 저장하여 일치도 검사를 수행할 수 있습니다.

각 Oracle 모델에 대한 키 문자열은 모델 목록 대화 상자의 모형 정보 열 아래에 표시됩니다. IBM SPSS Modeler 모델에 대한 키 문자열은 스트림에 배치될 때 IBM SPSS Modeler 모델의 서버 탭의 모델 키로 표시됩니다.

모델 너깃의 서버 탭에 있는 확인 단추를 사용하여 IBM SPSS Modeler 모델 및 Oracle 모델의 모델 키가 일치하는지 확인할 수 있습니다. 동일한 이름의 모델을 Oracle에서 발견할 수 없거나 모델 키가 일치하지 않으면 IBM SPSS Modeler 모델이 작성된 후에 Oracle 모델이 삭제되었거나 다시 작성된 것입니다.

### Oracle 모델 너깃 요약 탭

모델 너깃의 요약 탭에서는 모델 자체(분석), 모델에 사용된 필드(필드), 모델 작성 시 사용된 설정(작성 설정), 모델 학습(학습 요약)에 대한 정보를 표시합니다.

먼저 노드를 찾아볼 때 요약 탭 결과를 접습니다. 관심이 있는 결과를 보려면 항목 왼쪽에 있는 펼치기 제어를 사용하여 펼치거나 모두 펼치기 단추를 클릭하여 모든 결과를 표시합니다. 결과 보기와 완료한 경우 결과를 숨기려면 펼치기 제어를 사용하여 숨기려는 특정 결과를 접거나 모두 접기 단추를 클릭하여 모든 결과를 접으십시오.

**분석.** 특정 모델에 대한 정보를 표시합니다. 이 모델 너깃에 연결된 분석 노드를 실행한 경우에는 해당 분석의 정보도 이 섹션에 표시됩니다.

**필드.** 모델을 작성할 때 목표로 사용되는 필드와 입력을 나열합니다.

**작성 설정.** 모델을 작성할 때 사용되는 설정에 대한 정보를 포함합니다.

**학습 요약.** 모델 유형, 이를 작성하는 데 사용된 스트림, 이를 작성한 사용자, 작성 시점, 모델 작성 시 경과 시간을 표시합니다.

### Oracle 모델 너깃 설정 탭

모델 너깃의 설정 탭을 사용하면 스코어링 목적의 모델링 노드에서 특정 옵션의 설정을 대체할 수 있습니다.

Oracle 의사결정 트리

**오분류 비용 사용.** Oracle 의사결정 트리에서 오분류 비용을 사용할 것인지 여부를 결정합니다. 자세한 정보는 26 페이지의 『오분류 비용』의 내용을 참조하십시오.

**규칙 식별자.** 선택하면(체크하면) Oracle 의사결정 트리 모형에 규칙 식별자 열이 추가됩니다. 규칙 식별자는 특정 분할이 작성되는 트리 내의 노드를 식별합니다.

Oracle NMF

**모든 기능 표시.** 선택하면(체크하면) Oracle NMF 모델 내의 최고 기능에 대한 값만 표시되지 않고 모든 기능에 대한 기능 ID 및 신뢰도가 표시됩니다.

## Oracle 모델 나열

Oracle Data Mining 모델 나열 단추는 기존 데이터베이스 모델을 나열하고 모델을 제거할 수 있게 하는 대화 상자를 시작합니다. 이 대화 상자는 헬퍼 애플리케이션 대화 상자와 ODM 관련 노드에 대한 작성, 찾아보기 및 적용 대화 상자에서 시작할 수 있습니다.

각각의 모델에 대해 다음과 같은 정보가 표시됩니다.

- **모델 이름.** 목록을 정렬하는 데 사용되는 모델의 이름
- **모델 정보.** 작성 날짜/시간 및 목표 열 이름으로 구성된 모델 키 정보
- **모델 유형.** 이 모델을 작성한 알고리즘의 이름

## Oracle Data Miner

Oracle Data Miner는 Oracle 데이터 마이닝(ODM)에 대한 사용자 인터페이스이며 ODM에 대한 이전의 IBM SPSS Modeler 사용자 인터페이스를 대체합니다. Oracle Data Miner는 ODM 알고리즘을 적절히 활용하여 분석 가의 성공 비율을 높이도록 계획되었습니다. 이러한 목적은 여러 가지 방법으로 달성할 수 있습니다.

- 사용자가 데이터 준비 및 알고리즘 선택을 모두 다루는 방법을 적용하려면 더 많은 도움이 필요합니다. Oracle Data Miner는 적절한 방법을 통해 사용자를 안내하도록 데이터 마이닝 활동을 제공함으로써 이러한 요구를 충족합니다.
- Oracle Data Miner는 모델 작성 분야의 개선되고 확장된 휴리스틱을 포함하며 모델 및 변환 설정을 지정할 때 오류 발생 확률을 줄이기 위한 변환 마법사를 포함합니다.

Oracle Data Miner 연결 정의

1. Oracle Data Miner는 모든 Oracle 작성, 적용 노드 및 출력 대화 상자에서 **Oracle Data Miner** 시작을 통해 시작할 수 있습니다.



그림 2. *Oracle Data Miner* 시작 단추

2. Oracle Data Miner 외부 애플리케이션이 시작되기 전에 Oracle Data Miner 연결 편집 대화 상자가 표시됩니다. 단, 헬퍼 애플리케이션 옵션이 적절히 정의되어 있어야 합니다.

참고: 이 대화 상자는 정의된 연결 이름이 없는 경우에만 표시됩니다.

- 데이터 마이너 연결 이름을 제공하고 적절한 Oracle 10gR1 또는 10gR2 서버 정보를 입력하십시오. Oracle 서버가 IBM SPSS Modeler에서 지정된 서버와 동일해야 합니다.
3. Oracle Data Miner 연결 선택 대화 상자는 위 단계에서 정의한 연결 이름 중 사용할 이름을 지정하기 위한 옵션을 제공합니다.

Oracle Data Miner 요구 사항, 설치 및 사용법에 대한 자세한 정보는 Oracle 웹 사이트에서 [Oracle Data Miner](#)를 참조하십시오.

## 데이터 준비

모델링에서 Oracle Data Mining 알고리즘과 함께 제공된 Naive Bayes, 적응형 베이스 및 지원 벡터 머신을 사용하는 경우 두 가지 유형의 데이터 준비가 유용할 수 있습니다.

- **구간화**(연속형 숫자 범위 필드를 연속형 데이터를 승인할 수 없는 알고리즘에 대한 범주로 변환)
- **정규화**(비슷한 평균 및 표준 편차를 가지도록 숫자 범위에 적용된 변환)

구간화

IBM SPSS Modeler의 구간화 노드는 구간화 조작을 수행하는 데 필요한 다수의 기술을 제공합니다. 구간화 조작은 하나 이상의 필드에 적용할 수 있도록 정의됩니다. 데이터 세트에 대해 구간화 조작을 실행하면 임계값이 작성되고 IBM SPSS Modeler 파생 노드가 작성될 수 있습니다. 모델 작성 및 스코어링 전에 파생 조작을 SQL로 변환

하여 적용할 수 있습니다. 이 접근 방식에서는 구간화를 수행하지만 다중 모델 작업에서 구간화 사양을 재사용할 수 있게 하는 파생 노드와 모델 간 종속성을 작성할 수 있습니다.

#### 정규화

지원 벡터 머신 모델에 대한 입력으로 사용되는 연속형(숫자 범위) 필드는 모델 작성 전에 정규화해야 합니다. 회귀 모형의 경우에는 모델 출력에서 스코어를 재구성하기 위해 정규화를 되돌리기도 해야 합니다. SVM 모델 설정을 사용하면 **Z-스코어, 최소-최대 또는 없음**을 선택할 수 있습니다. 정규화 계수는 모델 작성 프로세스의 한 단계로 Oracle에 의해 구성되며 이 계수는 IBM SPSS Modeler에 업로드되고 모델과 함께 저장됩니다. 적용 시 이 계수는 IBM SPSS Modeler 파생 표현식으로 변환되고 데이터를 모델에 전달하기 전에 스코어링을 위해 데이터를 준비하는 데 사용됩니다. 이 경우 정규화는 모델링 작업과 밀접하게 연관되어 있습니다.

## Oracle 데이터 마이닝 예

IBM SPSS Modeler와 함께 ODM을 사용하는 방법을 나타내는 여러 표본 스트림이 포함됩니다. 이러한 스트림은 **|Demos|Database\_Modelling|Oracle Data Mining|** 아래의 IBM SPSS Modeler 설치 폴더에 있습니다.

참고: Demos 폴더는 Windows 시작 메뉴의 IBM SPSS Modeler 프로그램 그룹에서 액세스할 수 있습니다.

다음 표의 스트림은 Oracle 데이터 마이닝과 함께 제공되는 지원 벡터 머신(SVM)을 사용하여 데이터베이스 마이닝 프로세스의 예로서 순서대로 함께 사용될 수 있습니다.

표 4. 데이터베이스 마이닝 - 예 스트림	
스트림	설명
1_upload_data.str	플랫 파일에서 데이터베이스로 데이터를 정리하고 업로드하는 데 사용됩니다.
2_explore_data.str	IBM SPSS Modeler를 사용한 데이터 탐색의 예를 제공합니다.
3_build_model.str	데이터베이스의 원래 알고리즘을 사용하여 모델을 작성합니다.
4_evaluate_model.str	IBM SPSS Modeler를 사용하여 모형을 평가하는 예로 사용됩니다.
5_deploy_model.str	In-Database 스코어링에 대한 모델을 배포합니다.

참고: 예를 실행하려면 스트림이 순서대로 실행되어야 합니다. 또한 사용할 데이터베이스에 대한 유효한 데이터 소스를 참조하기 위해 각 스트림의 소스 및 모델링 노드가 업데이트되어야 합니다.

예제 스트림에서 사용된 데이터 세트는 신용카드 애플리케이션과 관련되며 분류 문제점에 범주형 예측변수와 연속형 예측변수의 혼합을 제공합니다. 이 데이터 세트에 대한 자세한 정보는 샘플 스트림과 동일한 폴더의 **crx.names** 파일을 참조하십시오.

이 데이터 세트는 UCI Machine Learning Repository <ftp://ftp.ics.uci.edu/pub/machine-learning-databases/credit-screening/>에서 사용 가능합니다.

## 예제 스트림: 데이터 업로드

첫 번째 예제 스트림인 **1\_upload\_data.str**은 플랫 파일의 데이터를 정리하고 Oracle로 업로드하는 데 사용됩니다.

Oracle Data Mining을 사용하려면 고유 ID 필드가 필요하므로 이 초기 스트림에서는 파생 노드를 통해 IBM SPSS Modeler @INDEX 함수를 사용하여 새 필드를 고유 값 1,2,3을 가진 ID라는 데이터 세트에 추가합니다.

채움 노드는 결측값 처리에 사용되며 텍스트 파일 **crx.data**로부터 읽어오는 비어 있는 필드를 **NULL** 값으로 바꿉니다.

## 예제 스트림: 데이터 탐색

두 번째 예제 스트림인 *2\_explore\_data.str*은 요약 통계 및 그래프를 포함하여 데이터의 일반적인 개요를 얻기 위해 데이터 검토 노드를 사용하는 방법을 보여주는 데 사용됩니다.

데이터 검토 보고서에서 그래프를 두 번 클릭하면 지정된 필드를 더 깊게 탐색할 수 있도록 자세한 그래프가 생성됩니다.

## 예제 스트림: 모델 작성

세 번째 예제 스트림인 *3\_build\_model.str*은 IBM SPSS Modeler에서 모델 작성을 보여줍니다. 데이터베이스 소스 노드(CREDIT로 레이블 지정됨)를 두 번 클릭하여 데이터 소스를 지정하십시오. 작성 설정을 지정하려면 작성 노드를 두 번 클릭하십시오(이 노드는 처음에는 CLASS로 레이블 지정되어 있지만 데이터 소스가 지정될 때 FIELD16으로 변경됨).

대화 상자의 모델 탭에서:

1. **ID**가 고유 필드로 선택되어 있는지 확인하십시오.
2. 선형이 커널 합수로 선택되어 있고 **Z-스코어**가 정규화 방법인지 확인하십시오.

## 예제 스트림: 모델 평가

네 번째 예제 스트림인 *4\_evaluate\_model.str*은 In-Database 모델링에 대해 IBM SPSS Modeler를 사용할 때의 장점을 보여줍니다. 모델을 실행하고 나면 해당 모델을 다시 데이터 스트림에 추가하고 IBM SPSS Modeler에서 제공된 여러 도구를 사용하여 해당 모델을 평가할 수 있습니다.

모델링 결과 보기

테이블 노드를 모델 너깃에 연결하여 결과를 탐색하십시오. **\$O-field16** 필드에는 각 케이스의 *field16*에 대한 예측값이 표시되고 **\$OC-field16** 필드에는 이 예측에 대한 신뢰도가 표시됩니다.

모델 결과 평가

분석 노드를 사용하여 각 예측 필드와 해당 대상 필드 간 일치의 패턴을 보여주는 일치 교차표를 작성할 수 있습니다. 분석 노드를 실행하여 결과를 확인하십시오.

평가 노드를 사용하여 모델에 의해 작성된 정확도 개선사항을 표시하도록 설계된 Gains 차트를 작성할 수 있습니다. 평가 노드를 실행하여 결과를 확인하십시오.

## 예제 스트림: 모델 배포

모델의 정확도에 만족한 경우에는 외부 애플리케이션과 함께 사용하거나 데이터베이스에 다시 게시하기 위해 해당 모델을 배포할 수 있습니다. 최종 예제 스트림인 *5\_deploy\_model.str*에서는 데이터를 CREDITDATA 테이블에서 읽어온 후 배포 솔루션이라는 발행자 노드를 사용하여 스코어링하여 CREDITSCORES 테이블에 게시합니다.

# 제 5 장 IBM Data Warehouse 및 IBM Netezza Analytics 를 사용한 데이터베이스 모델링

## IBM Data Warehouse 및 IBM Netezza Analytics가 포함된 SPSS Modeler

IBM SPSS Modeler는 IBM Data Warehouse 및 IBM Netezza<sup>®</sup> Analytics와의 통합을 지원하며 이를 통해 이러한 IBM 서버에서 고급 분석을 실행하는 기능을 제공합니다. 이 기능은 IBM SPSS Modeler 그래픽 사용자 인터페이스 및 워크플로우 지향 개발 환경을 통해 액세스할 수 있으며 이를 통해 IBM Netezza 또는 IBM Data Warehouse 환경에서 직접 데이터 마이닝 알고리즘을 실행할 수 있습니다.

SPSS Modeler는 **IBM Netezza Analytics**에서 제공하는 다음과 같은 알고리즘의 통합을 지원합니다.

- 의사결정 트리
- K-평균
- TwoStep
- Bayes Net
- Naive Bayes
- KNN
- 분열 군집
- PCA
- 회귀분석 트리
- 선형 회귀
- 시계열
- 일반화 선형

이러한 알고리즘에 대한 자세한 정보는 *IBM Netezza Analytics* 개발자 안내서 및 *IBM Netezza Analytics* 참조서를 참조하십시오.

SPSS Modeler는 **IBM Data Warehouse**에서 제공하는 다음과 같은 알고리즘의 통합을 지원합니다. 단, Bayes Net, 분열 군집 및 시계열은 지원되지 않습니다.

- 의사결정 트리
- K-평균
- TwoStep
- Naive Bayes
- KNN
- PCA
- 회귀분석 트리
- 선형 회귀
- 일반화 선형

참고: AIX는 지원되지 않습니다.

## 통합 요구사항

다음 조건은 IBM Netezza Analytics 또는 IBM Data Warehouse를 사용하여 In-Database 모델링을 수행하기 위한 전제조건입니다. 데이터베이스 관리자에게 문의하여 이 조건이 충족되는지 확인해야 할 수 있습니다.

- Windows 또는 UNIX(IBM Netezza ODBC 드라이버를 사용할 수 없는 zLinux 제외)의 IBM SPSS Modeler Server 설치에 대해 실행 중인 IBM SPSS Modeler
- IBM Netezza Analytics 패키지를 실행 중인 IBM Netezza Performance Server

**참고:** 필요한 Netezza Performance Server(NPS)의 최소 버전은 필요한 INZA의 버전에 따라 다르며 다음과 같습니다.

- NPS 6.0.0 P8 이상의 버전은 2.0 이전의 INZA 버전을 지원합니다.
- INZA 2.0 이상을 사용하려면 NPS 6.0.5 P5 이상이 필요합니다.

Netezza 일반화 선형 및 Netezza 시계열이 작동하려면 INZA 2.0 이상이 필요합니다. 기타 모든 Netezza In-Database 노드에는 INZA 1.1 이상이 필요합니다.

- IBM Netezza 데이터베이스에 연결하기 위한 ODBC 데이터 소스. 자세한 정보는 42 페이지의 [『통합 사용』](#)의 내용을 참조하십시오.
- IBM Data Warehouse 데이터베이스에 연결하기 위한 ODBC 데이터 소스
- IBM SPSS Modeler에서 사용으로 설정된 SQL 생성 및 최적화. 자세한 정보는 42 페이지의 [『통합 사용』](#)의 내용을 참조하십시오.

**참고:** 데이터베이스 모델링 및 SQL 최적화를 위해서는 IBM SPSS Modeler 컴퓨터에서 IBM SPSS Modeler Server 연결이 사용으로 설정되어 있어야 합니다. 이 설정이 사용 가능하면 데이터베이스 알고리즘에 액세스하고, SQL을 IBM SPSS Modeler에서 직접 푸시백하고, IBM SPSS Modeler Server에 액세스할 수 있습니다. 현재 라이센스 상태를 검증하려면 IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

#### 도움말 > 정보 > 추가 세부사항

연결성이 사용 가능으로 설정되면 라이센스 상태 탭에 **서버 사용 가능** 옵션이 표시됩니다.

## 통합 사용

IBM Netezza Analytics 또는 IBM Data Warehouse와의 통합을 사용으로 설정하려면 다음 단계를 수행해야 합니다.

- IBM Netezza Analytics 또는 IBM Data Warehouse 구성
- ODBC 소스 작성
- IBM SPSS Modeler에서 통합을 사용으로 설정
- IBM SPSS Modeler에서 SQL 생성 및 최적화를 사용으로 설정

이에 대해서는 아래의 절에 설명되어 있습니다.

## IBM Netezza Analytics 또는 IBM Data Warehouse 구성

IBM Netezza Analytics 또는 IBM Data Warehouse를 설치하고 구성하려면 해당 IBM 문서를 참조하십시오. 예를 들어, IBM Netezza Analytics의 경우 해당 제품과 함께 제공되는 *IBM Netezza Analytics* 설치 안내서를 참조하십시오. 이 안내서의 데이터베이스 권한 설정 절에는 IBM SPSS Modeler 스트림이 데이터베이스에 쓸 수 있도록 하기 위해 실행해야 하는 스크립트의 세부사항이 포함되어 있습니다.

**참고:** 교차표 계산에 의존하는 노드를 사용하려는 경우에는 CALL NZM..INITIALIZE();를 실행하여 교차표 엔진을 초기화해야 합니다. 그렇지 않으면 스토어드 프로시저 실행이 실패합니다. 초기화는 각 데이터베이스에 대한 일회성 설정 단계입니다.

## IBM Netezza Analytics에 대한 ODBC 소스 작성

IBM Netezza 데이터베이스와 IBM SPSS Modeler 사이에 연결을 사용하려면 ODBC 시스템 데이터 소스 이름(DSN)을 작성해야 합니다.

DSN을 작성하기 전에 ODBC 데이터 소스 및 드라이버와 IBM SPSS Modeler에서의 데이터베이스 지원에 대한 기본적인 이해가 필요합니다.

IBM SPSS Modeler Server에 대해 분산 모드에서 실행 중인 경우에는 서버 컴퓨터에서 DSN을 작성하십시오. 로컬(클라이언트) 모드에서 실행 중인 경우에는 클라이언트 컴퓨터에서 DSN을 작성하십시오.

## Windows 클라이언트

1. Netezza 클라이언트 CD에서 *nzodbcsetup.exe* 파일을 실행하여 설치 프로그램을 시작하십시오. 화면에 표시되는 지시사항에 따라 드라이버를 설치하십시오. 전체 지시사항을 보려면 IBM Netezza ODBC, JDBC 및 OLE DB 설치 및 구성 안내서를 참조하십시오.
  - a. DSN 작성.

참고: 메뉴 순서는 Windows 버전에 따라 다릅니다.

    - **Windows XP.** 시작 메뉴에서 제어판을 선택하십시오. 관리 도구를 두 번 클릭한 후 데이터 소스(**ODBC**)를 두 번 클릭하십시오.
    - **Windows Vista.** 시작 메뉴에서 제어판을 선택한 후 시스템 유지보수를 선택하십시오. 관리 도구를 두 번 클릭하고 데이터 소스(**ODBC**)를 선택한 후 열기를 클릭하십시오.
    - **Windows 7.** 시작 메뉴에서 제어판, 시스템 & 보안, 관리 도구를 차례로 선택하십시오. 데이터 소스 (**ODBC**)를 선택한 후 열기를 클릭하십시오.
  - b. 시스템 DSN 탭으로 이동한 후 추가를 클릭하십시오.
2. 목록에서 **NetezzaSQL**을 선택한 후 완료를 클릭하십시오.
3. Netezza ODBC 드라이버 설정 화면의 **DSN 옵션** 탭에서 선택한 데이터 소스 이름, IBM Netezza 서버의 호스트 이름 또는 IP 주소, 연결의 포트 번호, 사용 중인 IBM Netezza 인스턴스의 데이터베이스 및 데이터베이스 연결에 대한 사용자 이름 및 비밀번호 세부사항을 입력하십시오. 필드에 대한 설명을 보려면 도움말 단추를 클릭하십시오.
4. 연결 테스트 단추를 클릭하여 데이터베이스에 연결할 수 있는지 확인하십시오.
5. 성공적으로 연결되어 있으면 확인을 반복적으로 클릭하여 ODBC 데이터 소스 관리자 화면을 종료하십시오.

## Windows Server

Windows Server에 대한 프로시저는 Windows XP에 대한 클라이언트 프로시저와 동일합니다.

## UNIX 또는 Linux 서버

다음의 프로시저가 UNIX 또는 Linux 서버에 적용됩니다(IBM Netezza ODBC 드라이버를 사용할 수 없는 zLinux는 제외).

1. Netezza 클라이언트 CD/DVD에서 관련 <platform>cli.package.tar.gz 파일을 서버의 임시 위치에 복사하십시오.
2. **gunzip** 및 **tar** 명령을 사용하여 아카이브 컨텐츠를 추출하십시오.
3. 추출되는 *unpack* 스크립트에 실행 권한을 추가하십시오.
4. 스크립트를 실행하여 화면에 표시되는 프롬프트에 응답하십시오.
5. *modelersrv.sh* 파일을 편집하여 다음의 행을 포함하십시오.

```
. <SDAP Install Path>/odbc.sh  
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64  
NZ_ODBC_INI_PATH=<SDAP Install Path>; export NZ_ODBC_INI_PATH
```

예를 들어,

```
. /usr/IBM/SPSS/SDAP/odbc.sh  
LD_LIBRARY_PATH_64=$LD_LIBRARY_PATH:/usr/local/nz/lib64; export LD_LIBRARY_PATH_64  
NZ_ODBC_INI_PATH=/usr/IBM/SPSS/SDAP; export NZ_ODBC_INI_PATH
```

6. */usr/local/nz/lib64/odbc.ini* 파일을 찾아서 해당 컨텐츠를 SDAP과 함께 설치되는 *odbc.ini* 파일(\$ODBCINI 환경 변수에 의해 정의된 파일)에 복사하십시오.

참고: 64비트 Linux 시스템의 경우 **Driver** 모수는 32비트 드라이버를 잘못 참조합니다. 이전 단계의 odbc.ini 컨텐츠를 복사하는 경우에는 이 모수 내에서 적절하게 경로를 편집하십시오. 예를 들어, 다음과 같습니다.

```
/usr/local/nz/lib64/libnzodbc.so
```

7. Netezza DSN 정의를 편집하여 사용될 데이터베이스를 반영하십시오.
8. IBM SPSS Modeler Server를 다시 시작한 후 클라이언트에서 Netezza In-Database 마이닝 노드의 사용을 테스트하십시오.

## SPSS Modeler에서 통합 사용

1. IBM SPSS Modeler 기본 메뉴에서 다음을 선택하십시오.

도구 > 옵션 > 헬퍼 애플리케이션

2. **IBM Data Warehouse** 탭을 클릭하십시오.

**IBM Data Warehouse Analytics 통합 사용:** IBM SPSS Modeler 창의 맨 아래에 있는 데이터베이스 모델링 팔레트를 사용으로 설정(아직 표시되지 않은 경우)하고 IBM Data Warehouse 및 Netezza Data Mining 알고리즘에 대한 노드를 추가합니다.

**IBM Data Warehouse 연결:** 편집 단추를 클릭한 후 ODBC 소스를 작성할 때 설정한 IBM Data Warehouse 연결 문자열을 선택하십시오. 자세한 정보는 IBM Data Warehouse 관리 콘솔을 참조하십시오.

## SQL 생성 및 최적화 사용

매우 큰 데이터 세트에 대해 작업할 수도 있기 때문에 성능을 위해 IBM SPSS Modeler에서 SQL 생성 및 최적화 옵션을 사용으로 설정해야 합니다.

1. IBM SPSS Modeler 메뉴에서 다음을 선택하십시오.

도구 > 스트림 특성 > 옵션

2. 탐색 분할창에서 **최적화** 옵션을 클릭하십시오.

3. **SQL 생성** 옵션이 사용으로 설정되어 있는지 확인하십시오. 이 설정은 데이터베이스 모델링이 작동하기 위해 필수입니다.

4. **SQL 생성 최적화 및 기타 실행 최적화**를 선택하십시오(엄격하게 요구되지 않지만 최적화된 성능을 위해서는 강력하게 권장됨).

## IBM Netezza Analytics 및 IBM Data Warehouse를 사용한 모델 작성

각각의 지원되는 알고리즘에는 해당 모델링 노드가 있습니다. 노드 팔레트의 **데이터베이스 모델링** 탭에서 IBM Data Warehouse 및 IBM Netezza 모델링 노드에 액세스할 수 있습니다.

### 데이터 고려사항

데이터 소스에 있는 필드는 모델링 노드에 따라 다양한 데이터 유형의 변수를 포함할 수 있습니다. IBM SPSS Modeler에서는 데이터 유형이 측정 수준으로 알려져 있습니다. 모델링 노드의 필드 탭에서는 아이콘을 사용하여 해당 입력 및 대상 필드에 대해 허용되는 측정 수준 유형을 표시합니다.

**대상 필드** - 대상 필드는 예측하려는 값이 있는 필드입니다. 목표를 지정할 수 있는 경우 소스 데이터 필드 중 하나만 대상 필드로 선택할 수 있습니다.

**레코드 ID 필드** - 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. 소스 데이터가 ID 필드를 포함하지 않는 경우에는 다음 프로시저에 표시된 대로 파생 노드를 사용하여 이 필드를 작성할 수 있습니다.

1. 소스 노드를 선택하십시오.
2. 노드 팔레트의 필드 조작 탭에서 파생 노드를 두 번 클릭하십시오.
3. 캔버스에서 해당 아이콘을 두 번 클릭하여 파생 노드를 여십시오.

4. 파생 필드 필드에 예를 들어, ID를 입력하십시오.
5. 수식 필드에서 @INDEX를 입력한 후 확인을 클릭하십시오.
6. 파생 노드를 나머지 스트림에 연결하십시오.

**참고:** NUMERIC(18, 0) 데이터 유형을 사용하여 Netezza 데이터베이스에서 긴 숫자 데이터를 검색하는 경우 SPSS Modeler는 가져오는 동안 데이터를 반올림할 수 있습니다. 이 문제를 방지하기 위해 BIGINT 또는 NUMERIC(36, 0) 데이터 유형을 사용하여 데이터를 저장하십시오.

**참고:** 사용 가능한 필드 유형에 제한이 있으므로 측정 수준이 유형 없음이고 역할이 레코드 ID인 필드는 Netezza In-Database 모델링 노드(예: K-평균)에 표시되지 않습니다.

## 널값 처리

입력 데이터에 널값이 포함되어 있는 경우 일부 Netezza 노드를 사용하면 오류 메시지가 표시되거나 장기 실행 스트림이 발생할 수 있으므로 널값이 포함된 레코드는 제거하는 것이 좋습니다. 다음의 방법을 사용하십시오.

1. 선택 노드를 소스 노드에 연결하십시오.
2. 선택 노드의 모드 옵션을 삭제로 설정하십시오.
3. 조건 필드에서 다음을 입력하십시오.

```
@NULL(field1) [or @NULL(field2)[... or @NULL(fieldN)])
```

모든 입력 필드를 포함해야 합니다.

4. 선택 노드를 나머지 스트림에 연결하십시오.

## 모델 출력

IBM Data Warehouse 또는 Netezza 모델링 노드가 포함된 스트림은 실행될 때마다 약간 다른 결과를 생성할 수 있습니다. 이는 모델 작성 전에 데이터를 임시 테이블로 읽어오므로 노드가 소스 데이터를 읽는 순서가 항상 동일하지 않기 때문입니다. 하지만 이 영향에 의해 생성된 차이는 무시할 수 있습니다.

## 일반 주석

- IBM SPSS Collaboration and Deployment Services에서는 IBM Data Warehouse 또는 IBM Netezza 데이터 베이스 모델링 노드가 포함된 스트림을 사용하여 스코어링 구성을 작성할 수 없습니다.
- Data Warehouse 또는 Netezza 노드에서 작성된 모델에 대해서는 PMML 내보내기 또는 가져오기를 수행할 수 없습니다.

## 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여** 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.

**목표.** 예측에 대한 목표로 하나의 필드를 선택하십시오. 일반화 선형 모형의 경우 이 화면의 **시행** 필드도 참조하십시오.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## 서버 옵션

서버 탭에서 모델을 작성할 IBM Data Warehouse 데이터베이스를 지정합니다.

**IBM Data Warehouse 서버 세부사항.** 여기서는 모델에 사용할 데이터베이스에 대한 연결 세부사항을 지정합니다.

- **업스트림 연결 사용.** (기본값) 업스트림 노드(예: 데이터베이스 소스 노드)에 지정된 연결 세부사항을 사용합니다. 이 옵션은 모든 업스트림 노드가 SQL 푸시백을 사용할 수 있는 경우에만 작동합니다. 이 경우에는 SQL이 모든 업스트림 노드를 완전하게 구현하므로 데이터를 데이터베이스 밖으로 이동하지 않아도 됩니다.
- **데이터를 연결로 이동.** 여기에서 지정하는 데이터베이스로 데이터를 이동합니다. 이를 수행하면 데이터가 다른 IBM Data Warehouse 데이터베이스 또는 다른 벤더의 데이터베이스에 있거나 데이터가 플랫 파일인 경우에도 모델링이 작동할 수 있습니다. 또한, 노드가 SQL 푸시백을 수행하지 않아 데이터가 추출된 경우에는 여기에 지정된 데이터베이스로 데이터가 다시 이동합니다. 편집 단추를 클릭하여 연결을 찾아서 선택할 수 있습니다.



**경고:** IBM Netezza Analytics 및 IBM Data Warehouse는 일반적으로 매우 큰 데이터 세트와 함께 사용됩니다. 데이터베이스 사이에서 또는 데이터베이스 안팎으로 많은 양의 데이터를 전송하면 시간이 많이 걸릴 수 있으므로 가능하면 피해야 합니다.

**참고:** ODBC 데이터 소스 이름이 각 IBM SPSS Modeler 스트림에 효과적으로 임베드됩니다. 한 호스트에서 작성된 스트림이 다른 호스트에서 실행되는 경우 데이터 소스의 이름이 각 호스트에서 동일해야 합니다. 그렇지 않으면 각 소스 또는 모델링 노드의 서버 탭에서 다른 데이터 소스가 선택될 수 있습니다.

## 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 스코어링 옵션에 대한 기본값을 설정할 수 있습니다.

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**이름이 사용된 경우 기존 이름 바꾸기.** 이 확인 상자를 선택하면 동일한 이름을 가진 모든 기존 모델을 덮어씁니다.

**스코어링에 사용 가능.** 모델 너겟에 대한 대화 상자에 표시될 스코어링 옵션에 대한 기본값을 여기서 설정할 수 있습니다. 이 옵션에 대한 세부사항은 특정 너겟의 설정 탭에 대한 도움말 항목을 참조하십시오.

## 모델 관리

SPSS Modeler를 통해 IBM Netezza 또는 IBM Data Warehouse 모델을 작성하면 SPSS Modeler에서는 모델이 작성되고 IBM Data Warehouse 데이터베이스에서는 모델이 작성되거나 바뀝니다. 이 유형의 SPSS Modeler 모델은 데이터베이스 서버에 저장된 데이터베이스 모델의 컨텐츠를 참조합니다. SPSS Modeler는 SPSS Modeler 모델과 Netezza 또는 Data Warehouse 모델 모두에서 동일한 생성된 모델 키 문자열을 저장하여 일관성 검사를 수행할 수 있습니다.

각 Netezza 또는 Data Warehouse 모델에 대한 모델 이름은 데이터베이스 모델 나열 대화 상자의 모델 정보 열 아래에 표시됩니다. SPSS Modeler 모델에 대한 모델 이름은 SPSS Modeler 모델(스트림에 배치된 경우)의 서버 탭에서 모델 키로 표시됩니다.

확인 단추는 Netezza 또는 Data Warehouse 모델 및 SPSS Modeler 모델의 모델 키가 일치하는지 확인하는 데 사용할 수 있습니다. 동일한 이름의 모델을 Netezza 또는 Data Warehouse에서 찾을 수 없거나 모델 키가 일치하지 않으면 SPSS Modeler 모델이 작성된 후 Netezza 또는 Data Warehouse 모델이 삭제되었거나 다시 작성된 것입니다.

## 데이터베이스 모델 나열

SPSS Modeler는 IBM Data Warehouse에 저장된 모델을 나열하는 대화 상자를 제공하고 모델을 삭제할 수 있게 합니다. 이 대화 상자는 IBM 헬퍼 애플리케이션 대화 상자와 IBM Data Warehouse 및 IBM Netezza 데이터 마이닝 관련 노드에 대한 작성, 찾아보기 및 적용 대화 상자에서 액세스할 수 있습니다. 각각의 모델에 대해 다음과 같은 정보가 표시됩니다.

- 모델 이름(목록을 정렬하는 데 사용되는 모델의 이름)
- 소유자 이름
- 모델에서 사용되는 알고리즘
- 모델의 현재 상태(예: 완전)
- 모델이 작성된 날짜

## IBM Data WH 회귀분석 트리

회귀분석 트리는 케이스 표본을 반복적으로 분할하여 숫자 대상 필드의 값을 기반으로 동일한 종류의 서브세트를 파생하는 트리 기반 알고리즘입니다. 의사결정 트리와 마찬가지로, 회귀분석 트리는 트리의 리프가 충분히 작거나 충분히 균일한 서브세트에 해당되는 여러 서브세트로 데이터를 분해합니다. 분할은 목표 속성 값의 산포도를 줄이기 위해 선택합니다. 그러면 리프에서의 해당 평균 값을 사용하여 목표 속성 값을 상당히 잘 예측할 수 있습니다.

### IBM Data WH 회귀분석 트리 작성 옵션 - 트리 성장

트리 성장 및 트리 가지치기를 위한 작성 옵션을 설정할 수 있습니다.

다음 작성 옵션을 트리 성장에 사용할 수 있습니다.

**최대 트리 깊이.** 루트 노드 아래에서 트리가 확장될 수 있는 최대 수준 수, 즉 표본이 반복적으로 분할되는 횟수입니다. 기본값은 62이며, 이 값은 모델링을 위한 최대 트리 깊이입니다.

**참고:** 모델 너깃의 뷰어가 모델을 텍스트 형식으로 표현하는 경우 최대 12수준의 트리가 표시됩니다.

**분할 기준.** 이 옵션은 트리 분할 중지 시점을 제어합니다. 기본값을 사용하지 않으려면 **사용자 정의**를 클릭하고 값을 변경하십시오.

- **분할 평가 측도.** 이 클래스 평가 측도는 트리를 분할하기에 가장 적합한 위치를 평가합니다.

**참고:** 현재, 분산이 사용 가능한 유일한 옵션입니다.

- **분할할 최소 개선도.** 트리에서 새 분할이 작성되기 전에 제거되어야 하는 불순도의 최소 양입니다. 트리 작성의 목표는 유사한 출력 값을 가진 하위 그룹을 작성하여 각 노드 내의 불순도를 최소화하는 것입니다. 분기에 대한 최상의 분할이 분할 기준에 의해 지정된 수치 미만으로 불순도를 감소시키는 경우 분기가 분할되지 않습니다.
- **분할할 인스턴스의 최소수.** 분할할 수 있는 최소 레코드 수입니다. 분할되지 않은 레코드가 이 수 미만으로 남아 있는 경우, 추가 분할이 수행되지 않습니다. 이 필드를 사용하여 트리에서 작은 하위 그룹이 작성되는 것을 방지할 수 있습니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.
- **없음.** 모델을 스코어링하는데 필요한 통계량만 포함됩니다.

### IBM Data WH 트리 작성 옵션 - 트리 가지치기

가지치기 옵션을 사용하여 회귀분석 트리의 가지치기 기준을 지정할 수 있습니다. 가지치기의 목적은 새 데이터에 대한 예상 정확도를 향상시키지 않는 지나치게 확장된 하위 그룹을 제거하여 과적합 위험을 줄이기 위한 것입니다.

**가지치기 측도.** 가지치기 측도를 사용하면 트리에서 리프를 제거한 후에 모델의 추정 정확도가 허용 가능한 한계내로 유지될 수 있습니다. 다음 측도 중 하나를 선택할 수 있습니다.

- **mse.** 평균 제곱 오차 - (기본값) 맞춤선이 데이터 점에 어느 정도 가까운지 측정합니다.
- **r2.** R-제곱 - 회귀 모형에 의해 설명되는 종속변수에서 편차의 비율을 측정합니다.
- **Pearson.** Pearson 상관 계수 - 정규적으로 분포된 선형 종속 변수 간의 관계 강도를 측정합니다.

- **Spearman.** Spearman 상관 계수 - Pearson 상관에 따라 약해 보이지만 실제로는 강할 수 있는 비선형 관계를 발견합니다.

**가지치기를 위한 데이터.** 학습 데이터의 일부 또는 모두를 사용하여 새 데이터에 대한 예상 정확도를 추정할 수 있습니다. 또는 이 용도로 지정된 테이블에서 별도의 가지치기 데이터 세트를 사용할 수 있습니다.

- **모든 학습 데이터 사용.** 기본값인 이 옵션은 모든 훈련 데이터를 사용하여 모형 정확도를 추정합니다.
  - **가지치기를 위해 학습 데이터의 % 사용.** 가지치기 데이터에 대해 여기서 지정된 백분율을 사용하여 데이터를 두 개의 세트(학습에 대한 세트 하나와 가지치기에 대한 세트 하나)로 분할하려면 이 옵션을 사용하십시오.
- 난수 시드를 지정하여 스트림을 실행할 때마다 데이터가 동일한 방식으로 파티셔닝되도록 하려면 결과 복제를 선택하십시오. 가지치기에 사용되는 시드 필드에 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.
- **기준 테이블의 데이터 사용.** 모형 정확도 추정을 위한 별도의 가지치기 데이터 세트의 테이블 이름을 지정하십시오. 이 방법은 학습 데이터를 사용하는 것보다 신뢰성이 높은 것으로 간주됩니다. 하지만 이 옵션을 사용하면 학습 세트에서 데이터의 큰 서브세트가 제거되어 의사결정 트리의 품질이 저하될 수 있습니다.

## Netezza 분열 군집

분열 군집은 지정된 중지 포인트에 도달할 때까지 군집을 부군집으로 분열하기 위해 알고리즘이 반복적으로 실행되는 군집분석 방법입니다.

군집 구조는 모든 학습 인스턴스(레코드)를 포함하는 단일 군집으로 시작합니다. 알고리즘의 처음 반복은 데이터 세트를 두 개의 부군집으로 분열하고 후속 반복은 이러한 부군집을 더 작은 부군집으로 분열합니다. 중지 기준은 데이터 세트가 분열되는 최대 수준 수인 최대반복수 및 추가 파티셔닝에 대한 인스턴스의 최소 필요 수로 지정됩니다.

결과적으로 발생하는 계층적 군집 트리는 다음 예에서 보듯이 루트 군집에서 아래로 전파하여 인스턴스를 분류하는 데 사용할 수 있습니다.

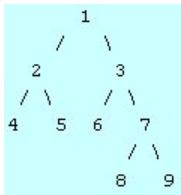


그림 3. 분열 군집 트리의 예

부군집 중심에서부터 인스턴스의 거리를 고려하여 각 수준에서 가장 많이 일치하는 부군집이 선택됩니다.

인스턴스가 -1(기본값)이라는 계층 수준이 적용되어 스코어링되는 경우, 리프가 음수로 지정되므로 스코어링이 리프 군집만 리턴합니다. 예에서는 군집 4, 5, 6, 8 또는 9 중 하나가 될 수 있습니다. 그러나 예를 들어, 계층 수준이 2로 설정되면 스코어링이 루트 군집 아래의 두 번째 수준에서 군집 중 하나, 즉 4, 5, 6 또는 7을 리턴합니다.

## Netezza 분열 군집 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.**

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## Netezza 분열 군집 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, 실행 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성옵션을 사용자 정의하려고 합니다.

**거리 측도.** 데이터 점 사이의 거리를 측정하는 데 사용할 방법입니다. 거리가 클수록 상이성도 커집니다. 옵션은 다음과 같습니다.

- **유클리디안.** (기본값) 두 점 사이의 거리는 두 점을 직선으로 결합하여 계산됩니다.
- **맨해튼.** 두 점 사이의 거리는 해당 좌표 간 절대 차이의 합계로 계산됩니다.
- **캔버라.** 맨해튼 거리와 비슷하지만 원점에 더 가까운 데이터 점에 더 민감합니다.
- **최대.** 두 점 사이의 거리가 좌표 차원을 따라 해당 차이 중 가장 큰 값으로 계산됩니다.

**최대 반복 수.** 동일한 프로세스를 여러 번 반복하여 작업하는 알고리즘입니다. 이 옵션을 사용하면 지정된 반복 수 이후에 모델 학습을 중지할 수 있습니다.

**군집 트리의 최대 깊이.** 데이터 세트가 소분열될 수 있는 최대 수준 수입니다.

**결과 복제.** 분석을 복제할 수 있도록 해주는 난수 시드를 설정하려면 이 선택란을 선택하십시오. 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.

**분할할 인스턴스의 최소수.** 분할할 수 있는 최소 레코드 수입니다. 분할되지 않은 레코드가 이 수 미만으로 남아 있는 경우, 추가 분할이 수행되지 않습니다. 이 필드를 사용하여 군집 트리에서 너무 작은 하위 그룹이 작성되는 것을 방지할 수 있습니다.

## IBM Data WH 일반화 선형

선형 회귀는 숫자 입력 필드 값에 기반하여 레코드를 분류하는 오랫동안 행해진 통계 기법입니다. 선형 회귀는 예측 및 실제 출력 값 사이의 차이를 최소화하는 직선 또는 곡선에 적합합니다. 선형 모형은 학습 및 모델 애플리케이션 들 다에서 단순성을 가지므로 광범위한 현실 세계의 현상을 모델링하는 데 유용합니다. 단, 선형 모형은 종속(목표)변수에서 정규 분포를 가정하고 독립(예측변수)변수가 종속변수에 선형 영향을 미친다고 가정합니다.

선형 회귀는 유용하나 위의 가정이 적용되지 않는 상황이 많이 있습니다. 예를 들어, 이산형 수의 곱 사이에서 소비자 선택을 모델링하는 경우, 종속변수가 다항분포를 이를 수 있습니다. 이와 유사하게 나이에 대한 수입을 모델링하는 경우, 일반적으로 나이가 증가하면 수입도 증가하나 둘 사이의 연결이 직선처럼 단순하지는 않습니다.

이러한 상황에 대해 일반화 선형 모형을 사용할 수 있습니다. 일반화 선형 모형은 적합한 함수라는 선택이 있는 경우에 지정된 연결함수를 사용하여 종속변수가 예측자 변수와 관련되도록 선형 회귀를 펼칩니다. 더욱이 이 모델을 사용하면 포아송과 같이 종속변수가 비정규 분포를 가질 수 있습니다.

알고리즘은 지정된 반복계산 수까지 반복적으로 최적 맞춤 모델을 찾습니다. 최적 맞춤을 계산하는 동안 종속변수의 예측값 및 실제 값의 차이의 제곱합으로 오차가 표시됩니다.

## IBM Data WH 일반화 선형 모델 필드 옵션

필드 탭에서, 이미 업스트림 노드에 정의된 필드 역할 설정을 사용할 것인지 여부를 선택하거나 필드에 수동으로 할당합니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드 또는 업스트림 소스 노드의 유형 탭의 역할 설정(예: 목표 또는 예측자)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.**

**목표.** 예측에 대한 목표로 하나의 필드를 선택하십시오.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다. 이 필드의 값은 고객 ID 번호 등과 같은 각 레코드에 대해 고유해야 합니다.

**인스턴스 가중치.** 인스턴스 가중치를 사용할 필드를 지정하십시오. 인스턴스 가중치는 입력 데이터의 행당 가중치입니다. 기본적으로, 모든 입력 레코드는 동일한 상대값 중요도를 갖고 있는 것으로 간주됩니다. 입력 레코드에 개별 가중치를 지정하여 중요도를 변경할 수 있습니다. 사용자가 지정하는 필드에는 입력 데이터의 각 행에 대한 숫자 가중치가 포함되어야 합니다.

**예측변수(입력).** 하나 이상의 입력 필드를 선택합니다. 이 동작은 유형 노드에서 필드 역할을 입력으로 설정하는 것과 유사합니다.

## IBM Data WH 일반화 선형 모델 옵션 - 일반

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 모델에 대한 여러 가지 설정, 연결함수, 입력 필드 상호작용(있는 경우에 한함) 및 스코어링 옵션에 대한 기본값을 작성할 수 있습니다.

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**필드 옵션.** 모델을 작성하기 위한 입력 필드의 역할을 지정할 수 있습니다.

**일반 설정.** 이러한 설정은 알고리즘에 대한 중지 기준과 관계가 있습니다.

- **반복 최대 수.** 알고리즘을 수행할 최대 반복 수입니다. 최소값은 1이며 기본값은 20입니다.
- **최대 오차(1e).** 알고리즘이 최적 맞춤 모델 찾기를 중지해야 하는 최대 오차 값(지수 표기법으로 표현)입니다. 최소값은 0이며 기본값은 -3으로, 1E-3 또는 0.001을 의미합니다.
- **무의미 오차 값 임계값(1e).** 그 이하에서 오류가 0의 값을 갖는 것으로 처리되는 값(지수 표기법으로 표현)입니다. 최소값은 -1이며 기본값은 -7로, 1E-7(또는 0.0000001) 아래의 오류 값은 무의미한 것으로 계수됨을 의미합니다.

**분포 설정.** 이러한 설정은 종속(목표) 변수의 분포와 연관됩니다.

- **반응 변수의 분포.** 분포 유형: 베르누이(기본값), 가우스, 포아송, 이항, 음이항, Wald(역가우스) 및 감마 중 하나입니다.
- **모수.** (포아송 또는 이항 분포 전용) 모수 지정 필드에서 다음 옵션 중 하나를 지정해야 합니다.
  - 데이터에서 자동으로 모수 추정값을 갖도록 하려면 기본값을 선택하십시오.
  - 분포 유사 우도의 최적화를 허용하려면 유사를 선택하십시오.
  - 모수값을 명시적으로 지정하려면 명시를 선택하십시오.

(이항 분포 전용) 이항 분포에 필요하므로 시행 필드로 사용할 입력 테이블 열을 지정해야 합니다. 이 열은 이항 분포에 대한 시행 수를 포함합니다.

(음이항 분포 전용) -1이라는 기본값을 사용하거나 다른 모수값을 지정할 수 있습니다.

**연결함수 설정.** 이러한 설정은 연결함수와 연관이 있으며 종속변수를 예측자 변수와 연관시킵니다.

- **연결함수.** 사용할 함수이며 항등, 역, Invnegative, Invsquare, Sqrt, 거듭제곱, 오즈 거듭제곱, 로그, C로그, 로그로그, C로그로그, 로짓(기본값), 프로빗, Gaussit, Cauchit, Canbinom, Cangeom, Cannegbinom 중 하나입니다.
- **모수.** (거듭제곱 또는 오즈 거듭제곱 연결함수 전용) 연결함수가 거듭제곱 또는 오즈 거듭제곱이면 모수값을 지정할 수 있습니다. 값을 지정하거나 기본값인 1을 사용하려면 선택하십시오.

## IBM Data WH 일반화 선형 모델 옵션 - 상호작용

상호작용 패널은 상호작용(입력 필드 사이의 승법 효과)을 지정하기 위한 옵션을 포함합니다.

**열 상호작용.** 입력 필드 사이의 상호작용을 지정하려면 이 확인 상자를 선택하십시오. 상호작용이 없으면 선택란을 그대로 두십시오.

소스 목록에서 하나 이상의 필드를 선택하고 상호작용은 목록으로 끌어 모델에 상호작용을 입력하십시오. 생성되는 상호작용의 유형은 선택을 놓는 핫스팟에 따라 다릅니다.

- **주.** 끌어 놓은 필드가 상호작용 목록 아래쪽에 별도의 주 상호작용으로 표시됩니다.
- **이원.** 끌어 놓은 필드의 모든 가능한 쌍이 상호작용 목록 아래쪽에 이원 상호작용으로 표시됩니다.

- **삼원.** 끌어 놓은 필드의 모든 가능한 세 개로 구성된 세트가 효과 목록 아래쪽에 삼원 상호작용으로 표시됩니다.
- \* 끌어 놓은 모든 필드의 조합은 상호작용 목록 아래쪽에 단일 상호작용으로 표시됩니다.

**절편 포함.** 이 모델에는 대개 절편이 포함됩니다. 데이터가 원점을 통하여 전달된다고 가정할 수 있는 경우 절편을 제외할 수 있습니다.

대화 상자 단추

표시 오른쪽의 단추를 사용하면 모델에서 사용되는 항을 변경할 수 있습니다.



그림 4. 삭제 단추

삭제할 항을 선택하고 삭제 단추를 클릭하여 모델에서 항목을 삭제할 수 있습니다.



그림 5. 다시 정렬 단추

다시 정렬할 항을 선택하고 위로 또는 아래로 화살표를 클릭하여 모델에서 항목을 다시 정렬할 수 있습니다.



그림 6. 사용자 정의 상호작용 단추

## 사용자 정의 항 추가

$n1*x1*x1*x1..$  양식으로 사용자 정의 상호작용을 지정할 수 있습니다. 필드 목록에서 필드를 선택하고 오른쪽 화살표 단추를 클릭하여 필드를 사용자 정의 항에 추가한 다음 곱\*을 클릭하고 다시 다음 필드를 선택한 다음 오른쪽 화살표 단추를 클릭하는 방법으로 계속 진행합니다. 사용자 정의 상호작용을 작성한 다음 항 추가를 클릭하여 상호작용 패널로 다시 돌아가십시오.

## IBM Data WH 일반화 선형 모델 옵션 - 스코어링 옵션

**스코어링에 사용 가능.** 모델 너겟에 대한 대화 상자에 표시될 스코어링 옵션에 대한 기본값을 여기서 설정할 수 있습니다. 자세한 정보는 [72 페이지의 『IBM Data WH 일반화 선형 모형 너겟 - 설정 탭』](#)의 내용을 참조하십시오.

- **입력 필드 포함.** 예측 및 모델 출력에 입력 필드를 표시하려면 이 선택란을 선택하십시오.

## IBM Data WH 의사결정 트리

의사결정 트리는 분류 모델을 표시하는 계층 구조입니다. 의사결정 트리 모형을 사용하면 분류 시스템을 개발하여 일련의 학습 데이터로부터 미래의 관측값을 예측 또는 분류할 수 있습니다. 분류는 분류의 분할 포인트를 표시하는 가지가 있는 트리 구조 형식을 사용합니다. 분할은 중지 포인트에 도달할 때까지 반복적으로 데이터를 하위 그룹으로 분류합니다. 중지 포인트의 트리 노드를 리프라고 합니다. 각 리프는 하위 그룹 또는 클래스의 멤버에 클래스 레이블로 알려진 레이블을 지정합니다.

## 인스턴스 가중치 및 클래스 가중치

기본적으로, 모든 입력 레코드와 클래스는 동일한 상대값 중요도를 갖고 있는 것으로 간주됩니다. 이러한 항목 중 하나 또는 둘 다의 멤버에 개별 가중치를 지정하여 이를 변경할 수 있습니다. 학습 데이터의 데이터 점이 범주 사이에 현실적으로 분배되지 않은 경우에 유용합니다. 가중치를 사용하면 데이터에서 제대로 표시되지 않는 범주

에 대해 보완할 수 있도록 모델을 편향시킬 수 있습니다. 목표 값에 대한 가중치를 늘리는 경우에는 해당 범주에 대한 올바른 예측의 백분율을 늘려야 합니다.

의사결정 트리 모델링 노드에서는 두 가지 유형의 가중치를 지정할 수 있습니다. **인스턴스 가중치**는 입력 데이터의 각 행에 가중치를 지정합니다. 다음 표에서 보듯이 가중치는 일반적으로 대부분의 케이스에 1.0으로 지정되고 대다수에 비해 중요도가 높거나 낮은 해당 케이스에 대해서만 더 높거나 낮은 값이 지정됩니다.

표 5. 인스턴스 가중치 예		
레코드 ID	목표	인스턴스 가중치
1	drugA	1.1
2	drugB	1.0
3	drugA	1.0
4	drugB	0.3

클래스 가중치는 다음 표에서 보듯이 대상 필드의 각 범주에 가중치를 지정합니다.

표 6. 클래스 가중치 예	
클래스	클래스 가중치
drugA	1.0
drugB	1.5

함께 곱하여 인스턴스 가중치로 사용하는 경우에 두 가지 유형의 가중치를 동시에 사용할 수 있습니다. 따라서 앞의 두 예가 함께 사용되는 경우에 다음 표에서 보듯이 알고리즘에 인스턴스 가중치가 사용될 수 있습니다.

표 7. 인스턴스 가중치 계산 예		
레코드 ID	계산	인스턴스 가중치
1	1.1*1.0	1.1
2	1.0*1.5	1.5
3	1.0*1.0	1.0
4	0.3*1.5	0.45

## Netezza 의사결정 트리 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)에서 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 수동으로 대상, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

목록의 모든 필드를 선택하려면 **모두** 단추를 클릭하거나 개별 측정 수준 단추를 클릭하여 이 측정 수준의 모든 필드를 선택하십시오.

**목표.** 예측에 대한 목표로 하나의 필드를 선택합니다.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다. 이 필드의 값은 각 레코드(고객 ID 번호 등)에 대해 고유해야 합니다.

**인스턴스 가중치.** 여기서 필드를 지정하면 대신 인스턴스 가중치(입력 데이터의 행당 가중치)를 사용하거나 기본 값인 클래스 가중치(대상 필드에 대한 범주당 가중치)를 사용할 수 있습니다. 여기서 지정하는 필드는 입력 데이

터의 각 행에 대한 숫자 가중치를 포함하는 필드여야 합니다. 자세한 정보는 [51 페이지의 『인스턴스 가중치 및 클래스 가중치』](#)의 내용을 참조하십시오.

**예측변수(입력).** 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 입력으로 설정하는 것과 유사합니다.

## IBM Data WH 의사결정 트리 작성 옵션

다음 작성 옵션을 트리 성장에 사용할 수 있습니다.

**확장 측도.** 이러한 옵션은 트리 성장이 측정되는 방법을 제어합니다.

- **불순도 측도.** 이 측도는 트리를 분할하기에 가장 적합한 위치를 평가합니다. 데이터 하위 그룹 또는 세그먼트에 서의 변동 측정치입니다. 낮은 불순도 측정치는 대부분의 멤버가 기준 또는 대상 필드에 대해 유사한 값을 갖는 그룹을 나타냅니다.

지원되는 측정은 **엔트로피** 및 **지니(Gini)**입니다. 이러한 측정은 분기에 대한 범주 소속 확률을 기반으로 합니다.

- **최대 트리 깊이.** 루트 노드 아래에서 트리가 확장될 수 있는 최대 수준 수, 즉 표본이 반복적으로 분할되는 횟수입니다. 이 특성의 기본값은 10이며 이 특성에 대해 설정할 수 있는 최대값은 62입니다.

**참고:** 모델 너깃의 뷰어가 모델을 텍스트 형식으로 표현하는 경우 최대 12수준의 트리가 표시됩니다.

**분할 기준.** 이 옵션은 트리 분할 중지 시점을 제어합니다.

- **분할할 최소 개선도.** 트리에서 새 분할이 작성되기 전에 제거되어야 하는 불순도의 최소 양입니다. 트리 작성의 목표는 유사한 출력 값을 가진 하위 그룹을 작성하여 각 노드 내의 불순도를 최소화하는 것입니다. 분기에 대한 최상의 분할이 분할 기준에 의해 지정된 수치 미만으로 불순도를 감소시키는 경우 분기가 분할되지 않습니다.
- **분할할 인스턴스의 최소수.** 분할할 수 있는 최소 레코드 수입니다. 분할되지 않은 레코드가 이 수 미만으로 남아 있는 경우, 추가 분할이 수행되지 않습니다. 이 필드를 사용하여 트리에서 작은 하위 그룹이 작성되는 것을 방지할 수 있습니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.

- **없음.** 모델을 스코어링하는데 필요한 통계량만 포함됩니다.

## IBM Data WH 의사결정 트리 노드 - 클래스 가중치

여기서는 개별 클래스에 가중치를 지정할 수 있습니다. 기본값은 모든 클래스에 동일한 가중치가 부여되도록 모든 클래스에 1 값을 지정하는 것입니다. 서로 다른 클래스 레이블에 대해 서로 다른 숫자 가중치를 지정함으로써 알고리즘이 그에 따라 특정 클래스의 학습 세트에 가중치를 주도록 합니다.

가중치를 변경하려면 **가중치** 열에서 가중치를 두 번 클릭하고 원하는 대로 변경하십시오.

**값.** 대상 필드의 가능한 값에서 파생된 클래스 레이블 세트입니다.

**가중치.** 특정 클래스에 지정할 가중치입니다. 클래스에 지정하는 가중치가 높을수록 모델이 해당 클래스에 대해 다른 클래스보다 더 민감하게 반응합니다.

클래스 가중치와 인스턴스 가중치를 조합하여 사용할 수 있습니다. 자세한 정보는 [51 페이지의 『인스턴스 가중치 및 클래스 가중치』](#)의 내용을 참조하십시오.

## IBM Data WH 의사결정 트리 노드 - 트리 가지치기

가지치기 옵션을 사용하여 의사결정 트리의 가지치기 기준을 지정할 수 있습니다. 가지치기의 목적은 새 데이터에 대한 예상 정확도를 향상시키지 않는 지나치게 확장된 하위 그룹을 제거하여 과적합 위험을 줄이기 위한 것입니다.

**가지치기 측도.** 기본 가지치기 측도인 정확도는 트리에서 리프를 제거한 후에 모델의 추정된 정확도가 허용 가능한 한계 내로 유지되도록 해줍니다. 가지치기를 적용하는 동안 클래스 가중치를 고려하는 경우에는 대신 가중 정확도를 사용하십시오.

**가지치기를 위한 데이터.** 학습 데이터의 일부 또는 모두를 사용하여 새 데이터에 대한 예상 정확도를 추정할 수 있습니다. 또는 이 용도로 지정된 테이블에서 별도의 가지치기 데이터 세트를 사용할 수 있습니다.

- **모든 학습 데이터 사용.** 기본값인 이 옵션은 모든 훈련 데이터를 사용하여 모형 정확도를 추정합니다.
- **가지치기를 위해 학습 데이터의 % 사용.** 가지치기 데이터에 대해 여기서 지정된 백분율을 사용하여 데이터를 두 개의 세트(학습에 대한 세트 하나와 가지치기에 대한 세트 하나)로 분할하려면 이 옵션을 사용하십시오.  
난수 시드를 지정하여 스트림을 실행할 때마다 데이터가 동일한 방식으로 파티셔닝되도록 하려면 결과 복제를 선택하십시오. 가지치기에 사용되는 시드 필드에 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.
- **기준 테이블의 데이터 사용.** 모형 정확도 추정을 위한 별도의 가지치기 데이터 세트의 테이블 이름을 지정하십시오. 이 방법은 학습 데이터를 사용하는 것보다 신뢰성이 높은 것으로 간주됩니다. 하지만 이 옵션을 사용하면 학습 세트에서 데이터의 큰 서브세트가 제거되어 의사결정 트리의 품질이 저하될 수 있습니다.

## IBM Data WH 선형 회귀

선형 모델은 목표와 하나 이상의 예측변수 사이의 선형 관계를 기반으로 연속형 목표를 예측합니다. 선형 관계를 직접 모델링하는 경우에 제한되기는 하나, 선형 회귀 모형은 비교적 단순하며 스코어링에 대해 쉽게 해석되는 수식을 제공합니다. 선형 모형은 신속하고 효율적이며 사용하기 쉽습니다. 단, 더 세분화된 회귀분석 알고리즘에 의해 생성된 모형에 비하면 적용성이 제한됩니다.

### IBM Data WH 선형 회귀 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, 실행 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

**방정식을 푸는 데 비정치값 분해를 사용.** 원래 교차표 대신 비정치값 분해 교차표를 사용하면 수치 오류에 대해 더 강력하다는 이점이 있으며 계산 속도가 빨라집니다.

**모델에 절편 포함.** 절편을 포함하면 솔루션의 전체 정확도가 높아집니다.

**모델 진단 계산.** 이 옵션을 사용하면 모델에서 수많은 진단이 계산됩니다. 결과는 행렬 또는 표에 저장됩니다. for later review. 진단에는 R 제곱, 잔차 제곱합, 분산 추정, 표준 편차,  $p$ -값 및  $t$ -값이 포함됩니다.

이러한 진단은 모델의 타당성 및 유용성과 관련됩니다. 선형성 가정을 충족하는지 확인하려면 기본 데이터에서 별도로 진단을 실행해야 합니다.

## IBM Data WH KNN

최근접 이웃 분석은 다른 케이스와의 유사성을 기준으로 케이스를 분류하는 방법입니다. 머신 훈련에서 이 분석 방법은 저장된 모든 패턴이나 케이스와 정확히 일치할 필요가 없는 데이터 패턴을 인식하는 방법으로 개발되었습니다. 유사한 케이스는 서로 가까이에 있고 유사하지 않은 케이스는 서로 멀리 떨어져 있습니다. 따라서 두 케이스 사이의 거리는 두 케이스의 상이성 측도가 됩니다.

서로 인접한 케이스를 "이웃"이라고 합니다. 새 케이스(검증용)가 있는 경우, 해당 모델에서 각 케이스와의 거리가 계산됩니다. 가장 유사한 케이스(최근접 이웃)의 분류가 기록되고 새 케이스가 최근접 이웃의 수가 가장 많은 범주에 배치됩니다.

탐색할 최근접 이웃 수를 지정할 수 있으며, 이 값을  $k$ 라고 합니다. 그림은 새 케이스가 두 개의 다른  $k$  값을 사용하여 분류되는 방법을 보여줍니다.  $k = 5$ 일 경우, 대부분의 최근접 이웃이 범주 1에 속하기 때문에 새 케이스는 범주 1에 위치합니다. 그러나  $k = 9$ 일 경우, 대부분의 최근접 이웃이 범주 0에 속하기 때문에 새 케이스는 범주 0에 위치합니다.

또한 최근접 이웃 분석은 연속적인 목표 값을 계산하는 데 사용할 수 있습니다. 이 경우, 가장 가까운 이웃의 평균 또는 중앙값 목표 값이 사용되어 새 케이스의 예측값을 가져옵니다.

## IBM Data WH KNN 모델 옵션 - 일반

모형 옵션 - 일반 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 또한 최근접 이웃 수를 계산하는 방법을 제어하는 옵션을 설정하고 강화된 성능 및 모델의 정확도에 대한 옵션을 설정 할 수 있습니다.

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

이웃

**거리 측도.** 데이터 점 사이의 거리를 측정하는 데 사용할 방법입니다. 거리가 클수록 상이성도 커집니다. 옵션은 다음과 같습니다.

- **유클리디안.** (기본값) 두 점 사이의 거리는 두 점을 직선으로 결합하여 계산됩니다.
- **맨해튼.** 두 점 사이의 거리는 해당 좌표 간 절대 차이의 합계로 계산됩니다.
- **캔버라.** 맨해튼 거리와 비슷하지만 원점에 더 가까운 데이터 점에 더 민감합니다.
- **최대.** 두 점 사이의 거리가 좌표 차원을 따라 해당 차이 중 가장 큰 값으로 계산됩니다.

**최근접 이웃 수(k).** 특정 케이스에 대한 최근접 이웃 수입니다. 많은 수의 이웃을 사용한다고 해서 반드시 더 정확한 모델을 얻을 수 있는 것은 아님에 유의하십시오.

$k$ 를 선택하면 과적합 방지("불량" 데이터의 경우 특히 중요할 수 있음)와 해결(비슷한 인스턴스에 대해 다양한 예측을 생성함) 사이의 균형이 제어됩니다. 일반적으로 1부터 수십까지의 일반적인 값 범위를 사용하여 각 데이터 세트에 대해  $k$ 의 값을 조정해야 합니다.

성능 및 정확성 강화

**거리를 계산하기 전에 측정 표준화.** 선택된 경우 이 옵션은 거리 값을 계산하기 전에 연속 입력 필드에 대한 측정을 표준화합니다.

**큰 데이터 세트의 성능 향상을 위해 코어 세트 사용.** 선택된 경우 이 옵션은 코어 세트 표본 추출을 사용하여 큰 데이터 세트가 관련된 경우 계산 속도를 높입니다.

## IBM Data WH KNN 모델 옵션 - 스코어링 옵션

모델 옵션 - 스코어링 옵션 탭에서 스코어링 옵션에 대한 기본값을 설정하고 개별 클래스에 상대 가중치를 지정 할 수 있습니다.

### 점수에 사용 가능

**입력 필드 포함.** 입력 필드가 기본적으로 스코어링에 포함되는지 여부를 지정합니다.

### 클래스 가중치

모델 작성에서 개별 클래스의 상대적 중요도를 변경하려면 이 옵션을 사용하십시오.

참고: 이 옵션은 분류에 KNN을 사용 중인 경우에만 사용 가능합니다. 회귀분석을 수행 중인 경우, 즉, 대상 필드 유형이 연속형인 경우에는 이 옵션을 사용할 수 없습니다.

기본값은 모든 클래스에 동일한 가중치가 부여되도록 모든 클래스에 1 값을 지정하는 것입니다. 서로 다른 클래스 레이블에 대해 서로 다른 숫자 가중치를 지정함으로써 알고리즘이 그에 따라 특정 클래스의 학습 세트에 가중치를 주도록 합니다.

가중치를 변경하려면 **가중치** 열에서 가중치를 두 번 클릭하고 원하는 대로 변경하십시오.

**값.** 대상 필드의 가능한 값에서 파생된 클래스 레이블 세트입니다.

**가중치.** 특정 클래스에 지정할 가중치입니다. 클래스에 지정하는 가중치가 높을수록 모델이 해당 클래스에 대해 다른 클래스보다 더 민감하게 반응합니다.

## IBM Data WH K-평균

K-평균 노드는 군집분석 방법을 제공하는 K-평균 알고리즘을 구현합니다. 이 노드를 사용하여 데이터 세트를 별개의 그룹으로 군집화할 수 있습니다.

이 알고리즘은 거리 메트릭(함수)을 기반으로 데이터 점 사이의 유사성을 측정하는 거리 기반 군집화 알고리즘입니다. 데이터 점은 사용되는 거리 메트릭에 따라 가장 가까운 군집에 지정됩니다.

각 훈련 인스턴스가 가장 가까운 군집에 지정되는 동일한 기본 프로세스를 여러 번 반복 수행함으로써 이 알고리즘이 작동합니다(지정된 거리 함수와 관련하여 해당 인스턴스 및 군집 중심에 적용됨). 그러면 모든 군집 중심이 특정 군집에 지정된 인스턴스의 평균 속성 값 벡터로서 다시 계산됩니다.

### IBM Data WH K-평균 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두** 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

### IBM Data WH K-평균 작성 옵션 탭

작성 옵션을 설정하여 고유의 목적에 맞게 모델 작성을 사용자 정의할 수 있습니다.

기본 옵션을 사용하여 모델을 작성하려면 **실행**을 클릭하십시오.

**거리 측도.** 이 매개변수는 데이터 점 사이의 거리를 측정하는 방법을 정의합니다. 거리가 멀수록 유사성이 떨어집니다. 다음 옵션 중 하나를 선택하십시오.

- 유클리디안.** 유클리디안 측도는 두 데이터 점 사이의 직선 거리입니다.
- 정규화 유클리디안.** 정규화 유클리디안 측도는 유클리디안 측도와 유사하지만 제곱 표준 편차에 의해 정규화됩니다. 유클리디안 측도와 달리, 정규화 유클리디안 측도는 척도 불변성(scale-invariant)을 가집니다.
- Mahalanobis의 거리.** Mahalanobis의 거리 측도는 입력 데이터의 상관계수를 고려하는 일반화 유클리디안 측도입니다. Mahalanobis의 거리 측도는 정규화 유클리디안 측도와 같이 규모 불변성을 갖습니다.
- Manhattan의 거리.** Manhattan의 거리 측도는 좌표 간의 절대 차이의 합으로 계산되는 두 데이터 점 사이의 거리입니다.
- Canberra의 거리.** Canberra의 거리 측도는 Manhattan의 거리 측도와 유사하나 원점에서 더 가까운 데이터 점에 대해 더 예민합니다.
- 최대값.** 최대값 측도는 좌표 차원을 따라 가장 큰 차이로 계산되는 두 데이터 점 사이의 거리입니다.

**군집 수.** 이 매개변수는 작성할 군집 수를 정의합니다.

**반복 최대 수.** 알고리즘이 동일한 프로세스를 여러 번 반복합니다. 이 매개변수는 모델 훈련이 중지하는 반복 수를 정의합니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- 모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- 열.** 열 관련 통계가 포함됩니다.

- 없음. 모델을 스코어링하는 데 필요한 통계량만 포함됩니다.

**결과 복제.** 분석을 복제하기 위한 난수 시드를 설정하려면 이 선택란을 선택하십시오. 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.

## IBM Data WH Naive Bayes

Naive Bayes는 분류 문제점을 위한 잘 알려진 알고리즘입니다. 이 모델은 제안된 모든 예측 변수를 서로 독립된 것으로 취급하므로 *naive*라는 이름이 사용됩니다. Naive Bayes는 속성과 목표 속성의 조합에 대한 조건부 확률을 계산하는 빠르고 확장 가능한 알고리즘입니다. 훈련 데이터로부터 독립적인 확률이 설정됩니다. 이 확률은 각 입력 변수에서 각각의 값 범주가 발생하는 경우 각 목표 클래스의 우도를 제공합니다.

## Netezza Bayes Net

베이지안 네트워크는 데이터 세트의 변수와 이 변수 사이의 확률적 또는 조건부 독립성을 표시하는 모델입니다. Netezza Bayes Net 노드를 사용하면 관측 및 기록한 증거를 상식적인 실세계 지식과 결합해서 곁보기에 링크되지 않은 속성을 사용하여 발생 우도를 설정함으로써 확률 모델을 작성할 수 있습니다.

### Netezza Bayes Net 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

이 노드의 경우, 스코어링에만 대상 필드가 필요하므로 이 탭에 표시되지 않습니다. 유형 노드, 이 노드의 모형 옵션 탭 또는 모델 너깃의 설정 탭에서 목표를 설정 또는 변경할 수 있습니다. 자세한 정보는 [66 페이지의 『Netezza Bayes Net 너깃 - 설정 탭』](#)의 내용을 참조하십시오.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두** 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

### Netezza Bayes Net 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, **실행** 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

**기준 지수.** 내부 관리를 더 쉽게 수행할 수 있도록 처음 속성(입력 필드)에 지정되는 숫자 식별자입니다.

**표본 크기.** 속성의 수가 너무 많아서 처리 시간이 너무 길어서 수용할 수 없는 경우에 사용하는 표본 크기입니다.

**실행 도중 추가 정보 표시.** 이 선택란을 선택하면(기본값), 메시지 대화 상자에 추가 진행률 정보가 표시됩니다.

## Netezza 시계열

시계열은 시간의 연속적이나 반드시 정기적이지는 않은 지점에서 측정되는 수치 데이터 값의 시퀀스입니다. 예를 들어, 매일 주가 또는 매주 판매 데이터 등이 있습니다. 추세 및 계절성(반복 패턴) 등의 동작을 강조하고 과거 이벤트로부터 미래 동작을 예측할 때 이러한 데이터 분석이 유용할 수 있습니다.

Netezza 시계열은 다음과 같은 시계열 알고리즘을 지원합니다.

- 스펙트럼 분석
- 지수평활
- 자기회귀 통합 이동 평균(ARIMA)

- 계절 추세 분해

이러한 알고리즘은 시계열을 추세 및 계절 성분으로 분해합니다. 그러면 예측에 사용할 수 있는 모델을 작성하기 위해 해당 성분을 분석할 수 있습니다.

**스펙트럼 분석**은 시계열에서 주기적 동작을 식별하는 데 사용됩니다. 다중 기본 주기성으로 구성된 시계열인 경우 또는 데이터에 상당한 양의 변량 잡음이 있는 경우, 스펙트럼 분석이 주기적 성분을 식별하는 가장 근접한 평균을 제공합니다. 이 방법은 계열을 시간 도메인에서 빈도 도메인 계열로 변환하여 주기적 동작의 빈도를 발견합니다.

**지수평활**은 향후 값을 예측하기 위해 이전 계열 관측의 가중된 값을 사용하는 시계열 분석 방법입니다. 지수평활과 함께 사용하면 지수 방법에서 지수평활의 영향력이 시간 경과에 따라 감소합니다. 이 방법은 덧셈, 추세 및 계절성을 고려하여 새 데이터가 들어오면 해당 예측을 조정하여 한 번에 하나의 포인트를 예측합니다.

**ARIMA** 모델은 지수평활 모델을 수행하는 모델링 추세 및 계절 성분에 대해 보다 정교한 방법을 제공합니다. 이 방법에는 차이 정도와 함께 자기회귀 및 이동 평균 순서를 명시적으로 지정하는 작업이 포함됩니다.

참고: 실제로 메일로 보내는 카탈로그 수 또는 회사 웹 페이지의 적중 수와 같은 예측할 계열의 동작을 설명하는데 도움이 될 수 있는 예측변수를 포함하려는 경우 ARIMA 모델이 가장 유용합니다. 지수평활 모델에서는 왜 원래대로 작동하는지 이유를 설명하지 않고도 시계열 동작을 설명합니다.

**계절 추세 분해**는 추세 분석을 수행한 다음 추세에 대한 기본 모양(2차 함수 등)을 수행하기 위해 시계열에서 정기적 동작을 제거합니다. 이러한 기본 모양에서는 잔차의 평균 제곱 오차(시계열의 맞춤값 및 관측값 사이의 차이)를 최소화하기 위한 모수의 수가 결정됩니다.

## Netezza 시계열에서 값의 보간법

보간법이란 시계열 데이터에서 결측값을 추정하고 삽입하는 프로세스입니다.

시계열의 구간이 규칙적이나 단순히 일부 값이 존재하지 않으면 선형 보간법을 사용하여 결측값을 추정할 수 있습니다. 매월 공항 터미널에 도착하는 승객의 계열을 생각해 보십시오.

표 8. 승객 터미널의 월별 도착	
월	승객
3	3,500,000
4	3,900,000
5	-
6	3,400,000
7	4,500,000
8	3,900,000
9	5,800,000
10	6,000,000

이 경우, 선형 보간법은 5월에 대한 결측값을 3,650,000(4월 및 6월 사이의 중심점)으로 추정합니다.

불규칙한 구간은 다르게 처리됩니다. 다음 온도 읽기 계열을 생각해 보십시오.

표 9. 온도 읽기		
날짜	시간	온도
2011-07-24	7:00	57
2011-07-24	14:00	75
2011-07-24	21:00	72
2011-07-25	7:15	59

표 9. 온도 읽기 (계속)

날짜	시간	온도
2011-07-25	14:00	77
2011-07-25	20:55	74
2011-07-27	7:00	60
2011-07-27	14:00	78
2011-07-27	22:00	74

사흘 동안 세 시점에서 온도를 읽었으나 시간이 다르며 그 중 일부만 공통됩니다. 또한 이틀만 연속적입니다.

이 상황은 두 가지 방법(통합 계산 또는 단계 크기 결정) 중 하나로 처리할 수 있습니다.

통합은 데이터에 대한 시맨틱 이해를 기반으로 하여 수식에 따라 계산되는 일일 통합입니다. 그러면 다음과 같은 데이터 세트가 작성될 수 있습니다.

표 10. 온도 읽기(통합)

날짜	시간	온도
2011-07-24	24:00	69
2011-07-25	24:00	71
2011-07-26	24:00	널
2011-07-27	24:00	72

또는 알고리즘이 계열을 구분 계열로 처리하고 적합한 단계 크기를 판별할 수 있습니다. 이 경우, 알고리즘에 의해 결정되는 단계 크기가 8시간이 되어 다음과 같은 결과가 발생할 수 있습니다.

표 11. 계산된 단계 크기로 온도 읽기

날짜	시간	온도
2011-07-24	6:00	
2011-07-24	14:00	75
2011-07-24	22:00	
2011-07-25	6:00	
2011-07-25	14:00	77
2011-07-25	22:00	
2011-07-26	6:00	
2011-07-26	14:00	
2011-07-26	22:00	
2011-07-27	6:00	
2011-07-27	14:00	78
2011-07-27	22:00	74

여기서는 원래 측정에 대해 네 개의 읽기만 해당되나 원래 계열의 다른 알려진 값의 도움을 받아 결측값이 보간법에 의해 다시 계산될 수 있습니다.

## Netezza 시계열 필드 옵션

필드 탭에서 소스 데이터의 입력 필드에 대한 역할을 지정하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**목표.** 예측에 대한 목표로 하나의 필드를 선택하십시오. 이 필드의 측정 수준이 연속형이어야 합니다.

**(예측변수) 시점.** (필수) 시계열의 날짜 또는 시간 값을 포함하는 입력 필드입니다. 필드의 측정 수준이 연속형 또는 범주형이어야 하며 데이터 저장 공간 유형이 날짜, 시간, 시간소인 또는 숫자여야 합니다. 여기서 지정하는 필드의 데이터 저장 공간 유형은 이 모델링 노드의 기타 탭의 일부 필드에 대한 입력 유형도 정의합니다.

**(예측변수) 시계열 ID(By).** 시계열 ID를 포함하는 필드로서, 입력이 둘 이상의 시계열을 포함하는 경우에 사용하십시오.

## Netezza 시계열 작성 옵션

두 가지 수준의 작성 옵션이 있습니다.

- 기본 - 알고리즘 선택, 보간법 및 사용할 시간 범위에 대한 설정입니다.
- 고급 - 시계열 분석에 대한 설정입니다.

이 절에서는 기본 옵션에 대해 설명합니다.

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, 실행 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

알고리즘

사용할 시계열 알고리즘과 관련된 설정입니다.

**알고리즘 이름.** 사용할 시계열 알고리즘을 선택하십시오. 사용 가능한 알고리즘은 스펙트럼 분석, 지수평활(기본 값), ARIMA 또는 계절 추세 분해입니다. 자세한 정보는 57 페이지의 『Netezza 시계열』의 내용을 참조하십시오.

**추세.** (지수평활 전용) 시계열이 추세를 나타내면 단순 지수평활이 잘 수행되지 않습니다. 지수가 있으면 알고리즘이 지수를 고려할 수 있도록 이 필드를 사용하여 지수를 지정하십시오.

- 결정된 시스템. (기본값) 시스템이 이 모수에 대한 최적 값을 찾으려고 시도합니다.
- 없음(N). 시계열이 추세를 나타내지 않습니다.
- 가법(A). 시간 경과에 따라 서서히 증가하는 추세입니다.
- 진폭감소 가법(DA). 결국 사라지는 가법 추세입니다.
- 승법(M). 시간 경과에 따라 증가하는 추세이며 일반적으로 점진적인 가법 추세보다 더 빠릅니다.
- 진폭감소 승법(DM). 결국 사라지는 승법 추세입니다.

**계절성.** (지수평활 전용) 시계열이 데이터에서 계절 패턴을 나타내는지 여부를 지정하려면 이 필드를 사용하십시오.

- 결정된 시스템. (기본값) 시스템이 이 모수에 대한 최적 값을 찾으려고 시도합니다.
- 없음(N). 시계열이 계절 패턴을 나타내지 않습니다.
- 가법(A). 계절 변동 패턴이 시간 경과에 따라 점진적 상승 추세를 나타냅니다.
- 승법(M). 가법 계절성과 동일하나 추가적으로 계절 변동의 진폭(높은 점 및 낮은 점 사이의 거리)이 변동의 전체 상승 추세에 비해 증가합니다.

**ARIMA에 대해 시스템 결정 설정 사용.** (ARIMA 전용) 시스템이 ARIMA 알고리즘에 대한 설정을 결정하도록 하려면 이 옵션을 선택하십시오.

**지정.** (ARIMA 전용) ARIMA 설정을 수동으로 지정하려면 이 옵션을 선택하고 단추를 클릭하십시오.

보간법

시계열 소스 데이터에 결측값이 있으면 값 추정을 삽입하는 방법을 선택하여 데이터 사이의 갭을 채우십시오. 자세한 정보는 58 페이지의 『Netezza 시계열에서 값의 보간법』의 내용을 참조하십시오.

- **선형.** 시계열의 구간이 규칙적이나 단순히 일부 값이 존재하지 않으면 이 방법을 선택하십시오.
- **지수 스플라인.** 알려진 데이터 점 값이 높은 비율로 증가하거나 감소하는 평활 곡선을 맞춥니다.
- **삼차 스플라인.** 결측값을 추정하기 위해 알려진 데이터 점으로 평활 곡선을 맞춥니다.

#### 시간 범위

시계열에서 전체 범위의 데이터를 사용할 것인지 또는 근접 데이터 서브세트를 사용하여 모델을 작성할 것인지 선택할 수 있습니다. 이러한 필드에 대한 유효한 입력은 필드 탭의 시점에 대해 지정된 필드의 데이터 저장 공간 유형에 의해 정의됩니다. 자세한 정보는 60 페이지의 [『Netezza 시계열 필드 옵션』](#)의 내용을 참조하십시오.

- **데이터에서 사용 가능한 가장 이른 시간 및 가장 최근의 시간 사용.** 전체 범위의 시계열 데이터를 사용하려면 이 옵션을 선택하십시오.
- **시간 창 지정.** 시계열의 일부만을 사용하려면 이 옵션을 선택하십시오. 경계를 지정하려면 **가장 이른 시간(시작)** 및 **최근의 시간(종료)** 필드를 사용하십시오.

## ARIMA 구조

ARIMA 모델에서 다양한 비계절 및 계절 성분의 값을 지정하십시오. 각 케이스에서 연산자를 =(같음) 또는 <=(이하)로 설정한 다음 인접 필드에서 값을 지정하십시오. 값은 차수를 지정하는 음이 아닌 정수여야 합니다.

**비계절.** 모델의 다양한 비계절 성분의 값입니다.

- **자기상관 차수(p).** 모델의 자기회귀 차수 수입니다. 자기회귀 차수는 계열의 이전 값 중 현재 값 예측에 사용될 값을 지정합니다. 예를 들어, 자기회귀 차수 2는 과거 2개 시간 주기의 계열 값을 현재 값 예측에 사용하도록 지정합니다.
- **파생(d).** 모델을 추정하기 전 계열에 적용할 차이 차수를 지정합니다. 추세가 존재하며(추세가 있는 계열은 일반적으로 비정상이며 ARIMA 모델링은 정상성을 가정) 해당 효과 제거를 위해 사용되는 경우 차이가 필요합니다. 차이 차수는 계열 추세 수준에 해당합니다. 1차 차이는 선형 추세, 2차 차이는 2차 추세 등을 나타냅니다.
- **이동 평균(q).** 모델의 이동 평균 차수 수입니다. 이동 평균 차수는 이전 값에 대한 계열 평균 편차를 사용하여 현재 값을 예측하는 방법을 지정합니다. 예를 들어, 이동 평균 차수 1과 2는 계열의 현재 값을 예측하는 경우 지난 2개 시간 주기 각각의 계열 평균값 편차를 고려하도록 지정합니다.

**계절.** 계절적 자기상관(SP), 파생(SD) 및 이동 평균(SQ) 성분이 비계절 상대로 동일한 역할을 수행합니다. 그러나 계절 차수의 경우 현재 계열 값이 한 개 이상의 계절 주기에 의해 구분된 이전 계열 값의 영향을 받습니다. 예를 들어, 월별 데이터(계절 주기 12)의 경우 계절 차수 1은 현재 계열 값이 현재 계열 이전의 계열 값 12 주기의 영향을 받는다는 것을 의미합니다. 월별 데이터의 경우 계절 차수 1은 비계절 차수 12를 지정하는 것과 동일합니다.

계절 설정은 데이터에서 계절성이 발견된 경우 또는 고급 탭에서 주기 설정을 지정한 경우에만 고려됩니다.

## Netezza 시계열 작성 옵션 - 고급

고급 설정을 사용하여 시계열 분석에 대한 옵션을 지정할 수 있습니다.

**모델 작성 옵션에 대해 시스템 결정 설정 사용.** 시스템이 고급 설정을 결정하도록 하려면 이 옵션을 선택하십시오.

**지정.** 고급 옵션을 수동으로 지정하려면 이 옵션을 선택하십시오. 알고리즘이 스펙트럼 분석이면 이 옵션을 사용 할 수 없습니다.

- **주/주기 단위.** 시계열의 몇 가지 공정특성 변수 작동이 반복된 후의 주기입니다. 예를 들어, 주말 세일 시계열에 대해 주기로 1을 지정하고 단위로 주를 지정할 수 있습니다. 주기는 음수가 아닌 정수여야 하며 주기 단위는 밀리초, 초, 분, 시, 일, 주, 분기 또는 년 중 하나여야 합니다. 주기가 설정되지 않았거나 시간 유형이 숫자가 아니면 주기 단위를 설정하지 마십시오. 단, 주기를 지정한 경우에는 반드시 주기 단위를 지정해야 합니다.

**시계열 분석 설정.** 특정 시점까지 또는 특정 시점에 예측값을 작성하도록 선택할 수 있습니다. 이러한 필드에 대한 유효한 입력은 필드 탭의 시점에 대해 지정된 필드의 데이터 저장 공간 유형에 의해 정의됩니다. 자세한 정보는 60 페이지의 [『Netezza 시계열 필드 옵션』](#)의 내용을 참조하십시오.

- **예측 범위.** 시계열 분석의 끝점만 지정하려면 이 옵션을 선택하십시오. 이 시점까지의 예측값이 작성됩니다.
- **예측 시간.** 예측값을 작성할 시점을 하나 이상 지정하려면 이 옵션을 선택하십시오. 시점 테이블에 새 행을 추가하려면 추가를 클릭하십시오. 행을 삭제하려면 행을 선택하고 삭제를 클릭하십시오.

## Netezza 시계열 모델 옵션

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다. 모델 출력 옵션에 대한 기본값을 설정할 수 있습니다.

**모델 이름.** 목표 또는 ID 필드(또는 이와 같은 필드가 지정되지 않은 경우 모델 유형)를 기준으로 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**스코어링에 사용 가능.** 모델 너겟에 대한 대화 상자에 표시될 스코어링 옵션에 대한 기본값을 여기서 설정할 수 있습니다.

• **결과에 히스토리 값 포함** 기본적으로 모델 출력은 히스토리 데이터 값(예측 작성에 사용된 값)을 포함하지 않습니다. 해당 값을 포함하려면 이 선택란을 선택하십시오.

• **결과에 보간값 포함.** 결과에 히스토리 값을 포함하도록 선택한 경우, 보간값이 있으면 보간값도 포함하려면 이 선택란을 선택하십시오. 보간법은 히스토리 데이터에 대해서만 적용되므로 **결과에 히스토리 값 포함**을 선택하지 않은 경우에는 이 선택란을 사용할 수 없습니다. 자세한 정보는 [58 페이지의 『Netezza 시계열에서 값의 보간법』](#)의 내용을 참조하십시오.

## IBM Data WH TwoStep

TwoStep 노드는 큰 데이터 세트에 대해 데이터를 군집시키는 TwoStep 알고리즘을 구현합니다.

이 노드를 사용하면 사용 가능한 자원(메모리 및 시간 제약조건 등)을 고려하면서 데이터를 군집화할 수 있습니다.

TwoStep 알고리즘은 다음과 같은 방법으로 데이터를 군집화하는 데이터베이스 마이닝 알고리즘입니다.

1. 군집화 기능(CF) 트리가 작성됩니다. 균형도가 높은 이 트리는 유사한 입력 레코드가 동일한 트리 노드의 일부가 되는 계층별 군집화를 위한 군집화 기능을 저장합니다.
2. CF 트리의 리프는 인메모리로 계층적으로 군집화되어 최종 군집 결과를 생성합니다. 최상의 군집 수는 자동으로 결정됩니다. 최대 군집 수를 지정하면 지정된 한계 내에서 최상의 군집 수가 결정됩니다.
3. 군집 결과는 K-평균 알고리즘과 유사한 알고리즘이 데이터에 적용되는 두 번째 단계에서 세분화됩니다.

## IBM Data WH TwoStep 필드 옵션

필드 옵션을 설정하여 업스트림 노드에 정의된 필드 역할 설정을 사용하도록 지정할 수 있습니다. 수동으로 필드 할당을 수행할 수도 있습니다.

**항목 선택.** 업스트림 소스 노드의 유형 탭 또는 업스트림 유형 노드의 역할 설정을 사용하려면 이 옵션을 선택하십시오. 역할 설정은, 예를 들면, 목표 및 예측변수입니다.

**사용자 정의 필드 할당 사용.** 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표를 사용하여 수동으로 이 목록의 항목을 화면의 오른쪽에 있는 역할 필드에 지정하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## IBM Data WH TwoStep 작성 옵션

작성 옵션을 설정하여 고유의 목적에 맞게 모델 작성을 사용자 정의할 수 있습니다.

기본 옵션을 사용하여 모델을 작성하려면 **실행**을 클릭하십시오.

**거리 측도.** 이 매개변수는 데이터 점 사이의 거리를 측정하는 방법을 정의합니다. 거리가 멀수록 유사성이 떨어집니다. 옵션은 다음과 같습니다.

- **로그-우도.** 우도 측도는 변수에 확률 분포를 둡니다. 연속형 변수는 정규 분포로, 범주형 변수는 다항분포로 가정됩니다. 모든 변수를 독립변수로 가정합니다.
- **유클리디안.** 유클리디안 측도는 두 데이터 점 사이의 직선 거리입니다.

• **정규화 유클리디안.** 정규화 유클리디안 측도는 유클리디안 측도와 유사하지만 제곱 표준 편차에 의해 정규화 됩니다. 유클리디안 측도와 달리, 정규화 유클리디안 측도는 척도 불변성(scale-invariant)을 가집니다.

군집 수. 이 매개변수는 작성할 군집 수를 정의합니다. 옵션은 다음과 같습니다.

- **자동으로 군집 수 계산.** 군집 수가 자동으로 계산됩니다. 최대값 필드에 군집 최대 수를 지정할 수 있습니다.
- **군집 수 지정.** 작성할 군집 수를 지정하십시오.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 옵션은 다음과 같습니다.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 없음을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.

- **없음.** 모델을 스코어링하는 데 필요한 통계만 포함됩니다.

**결과 복제.** 분석을 복제하기 위한 난수 시드를 설정하려면 이 선택란을 선택하십시오. 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.

## IBM Data WH PCA

비선형 주성분분석(PRINCALS)(PCA)는 데이터의 복잡도를 줄이기 위해 디자인된 강력한 데이터 축소 기법입니다. PCA는 전체 필드 세트에서 최상의 분산 캡처 작업을 수행하는 입력 필드의 선형 조합을 찾습니다. 이 때 성분은 서로 직교하며 상관분석되지 않습니다. 목표는 원래 입력 필드 세트의 정보를 효과적으로 요약하는 소수의 파생된 필드(주성분)를 찾는 것입니다.

**참고:** 소문자 필드 이름을 사용할 경우 모델을 스코어링할 때 오류가 발생할 수 있습니다. 이것은 알려진 Db2 Data Warehouse 결함으로, 임시 해결책은 스코어링하기 전에 모든 필드의 이름을 대문자로 변경하는 것입니다.

## IBM Data WH PCA 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.**

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## IBM Data WH PCA 작성 옵션

작성 옵션 탭은 모델을 작성하기 위한 모든 옵션을 설정하는 위치입니다. 물론, 실행 단추를 클릭하여 모든 기본 옵션으로 모델을 작성할 수도 있지만, 보통 사용자는 고유한 목적을 위해 작성을 사용자 정의하려고 합니다.

**PCA를 계산하기 전에 데이터 가운데 맞춤.** 이 옵션을 선택하면(기본값), 분석 전에 데이터 가운데 맞춤("평균 뺄셈"이라고도 함)을 수행합니다. 데이터 가운데 맞춤은 처음 주성분이 최대 분산의 방향을 설명하도록 하기 위해 꼭 필요합니다. 그렇지 않으면 성분이 데이터의 평균에 더 가까이 대응할 수 있습니다. 데이터가 이미 이 방법으로 준비된 경우에만 일반적으로 성능 향상을 위해 이 옵션을 선택 취소합니다.

**PCA를 계산하기 전에 데이터 배율 조정.** 이 옵션은 분석 전에 데이터 배율 조정을 수행합니다. 그렇게 하면 다른 변수가 다른 장치에서 측정될 때 분석을 덜 임의적으로 만들 수 있습니다. 가장 단순한 양식의 데이터 배율 조정은 각 변수를 표준 편차로 나누는 것입니다.

**PCA를 계산하기 위해 덜 정확하지만 더 빠른 방법 사용.** 이 옵션을 사용하면 알고리즘이 덜 정확하나 더 빠른 방법(forceEigensolve)을 사용하여 주성분을 찾습니다.

## IBM Data WH 및 Netezza 모델 관리

IBM Data Warehouse 및 IBM Netezza Analytics 모델은 다른 IBM SPSS Modeler 모델과 동일한 방식으로 캔버스 및 모델 팔레트에 추가되며 거의 동일한 방식으로 사용할 수 있습니다. 하지만 IBM SPSS Modeler에서 작성된 IBM Data Warehouse 또는 IBM Netezza Analytics 모델이 각각 실제로 데이터베이스 서버에 저장된 모델을 참조하는 경우에는 몇 가지 중요한 차이가 있습니다. 따라서 스트림이 올바르게 작동하려면 모델이 작성된 데이터베이스에 연결해야 하며 외부 프로세스에 의해 모델 테이블이 변경되지 않아야 합니다.

## IBM Data Warehouse 및 IBM Netezza Analytics 모델 스코어링

모델은 금색 모델 너깃 아이콘에 의해 캔버스에 표시됩니다. 너깃의 주 용도는 데이터를 스코어링하여 예측을 생성하거나 모델 특성의 추가 분석을 허용하는 것입니다. 나중에 이 절에서 설명되는 대로 테이블 노드를 너깃에 연결하고 스트림의 해당 분기를 실행하여 표시될 수 있는 하나 이상의 추가 데이터 필드 형식으로 스코어가 추가됩니다. 의사결정 트리 또는 회귀분석 트리에 대한 대화 상자 등의 일부 너깃 대화 상자에는 모델의 시각적 표시를 제공하는 모델 탭이 추가로 제공됩니다.

추가 필드는 대상 필드의 이름에 추가된 \$<id>- 접두부에 의해 구별됩니다. 여기서 <id>는 모델에 따라 다르며 추가 중인 정보의 유형을 식별합니다. 각각의 모델 너깃에 대한 주제에 다양한 식별자가 설명되어 있습니다.

스코어를 보려면 다음의 단계를 완료하십시오.

1. 테이블 노드를 모델 너깃에 연결하십시오.
2. 테이블 노드를 여십시오.
3. 실행을 클릭하십시오.
4. 테이블 출력 창의 오른쪽으로 스크롤하여 추가 필드 및 해당 스코어를 보십시오.

## IBM Data WH 및 Netezza 모델 너깃 서버 탭

서버 탭에서는 모델 스코어링을 위한 서버 옵션을 설정할 수 있습니다. 업스트림으로 지정된 서버 연결을 계속 사용하거나 여기서 지정하는 다른 데이터베이스로 데이터를 이동할 수 있습니다.

**IBM Data Warehouse 서버 세부사항.** 여기서는 모델에 사용할 데이터베이스에 대한 연결 세부사항을 지정합니다.

- **업스트림 연결 사용.** (기본값) 업스트림 노드(예: 데이터베이스 소스 노드)에 지정된 연결 세부사항을 사용합니다. 이 옵션은 모든 업스트림 노드가 SQL 푸시백을 사용할 수 있는 경우에만 작동합니다. 이 경우에는 SQL이 모든 업스트림 노드를 완전하게 구현하므로 데이터를 데이터베이스 밖으로 이동하지 않아도 됩니다.
- **데이터를 연결로 이동.** 여기에서 지정하는 데이터베이스로 데이터를 이동합니다. 이를 수행하면 데이터가 다른 IBM Data Warehouse 데이터베이스 또는 다른 벤더의 데이터베이스에 있거나 데이터가 플랫 파일인 경우에도 모델링이 작동할 수 있습니다. 또한, 노드가 SQL 푸시백을 수행하지 않아 데이터가 추출된 경우에는 여기에 지정된 데이터베이스로 데이터가 다시 이동합니다. 편집 단추를 클릭하여 연결을 찾아서 선택할 수 있습니다.



**경고:** IBM Netezza Analytics 및 IBM Data Warehouse는 일반적으로 매우 큰 데이터 세트와 함께 사용됩니다. 데이터베이스 사이에서 또는 데이터베이스 안팎으로 많은 양의 데이터를 전송하면 시간이 많이 걸릴 수 있으므로 가능하면 피해야 합니다.

**모델 이름.** 모델의 이름입니다. 이 이름은 정보용으로만 표시되며 여기서 변경할 수 없습니다.

## IBM Data WH 의사결정 트리 모델 너깃

의사결정 트리 모형 너깃은 모델링 작업의 출력을 표시하며 사용자가 모델 스코어링에 대한 일부 옵션을 설정할 수 있도록 해줍니다.

의사결정 트리 모형 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 12. 의사결정 트리의 모델 스코어링 필드	
추가되는 필드의 이름	의미
\$I-target_name	현재 레코드의 예측값입니다.

모델링 노드 또는 모델 너깃에서 스코어 레코드에 할당된 클래스의 확률 계산 옵션을 선택하고 스트림을 실행한 경우 추가적 필드가 추가됩니다.

표 13. 의사결정 트리에 대한 모델 스코어링 필드 - 추가적	
추가되는 필드의 이름	의미
\$IP-target_name	예측 신뢰도(0.0 - 1.0)입니다.

## IBM Data WH 의사결정 트리 너깃 - 모델 탭

모델 탭은 의사결정 트리 모형의 예측변수 중요도를 그래픽 형식으로 표시합니다. 막대의 길이는 예측변수의 중요도를 나타냅니다.

**참고:** IBM Netezza Analytics 2.x 이전 버전으로 작업 중인 경우에는 의사결정 트리 모형의 내용이 텍스트 형식으로만 표시됩니다.

이러한 버전에 대해서는 다음 정보가 표시됩니다.

- 노드 또는 리프에 해당하는 텍스트의 각 줄
- 트리 수준을 반영하는 들여쓰기
- 노드의 경우, 분할 조건이 표시됩니다.
- 리프의 경우, 지정된 클래스 레이블이 표시됩니다.

## IBM Data WH 의사결정 트리 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함:** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**스코어링 레코드에 대해 지정된 클래스의 확률 계산.** (의사결정 트리 및 Naive Bayes 전용) 선택된 경우 이 옵션은 추가 모델링 필드에 신뢰도(즉, 확률) 필드 및 예측 필드가 포함됨을 의미합니다. 이 선택란을 선택 취소하면 예측 필드만 생성됩니다.

**결정적 입력 데이터 사용.** 이 옵션을 선택하면 동일한 보기의 다중 패스를 실행하는 Netezza 알고리즘이 각 패스에 대해 동일한 데이터 세트를 사용합니다. 비결정적 데이터가 사용되고 있음을 표시하기 위해 이 선택란을 지우면 파티션 노드에 의해 생성되는 것과 같이 처리에 필요한 데이터 출력을 보유하기 위해 임시 테이블이 작성되고 모델이 작성된 후에 이 테이블이 삭제됩니다.

## IBM Data WH 의사결정 트리 너깃 - 뷰어 탭

뷰어 탭은 SPSS Modeler가 해당 의사결정 트리 모형의 트리 프리젠테이션을 표시하는 것과 동일한 방식으로 트리 모형의 트리 프리젠테이션을 표시합니다.

**참고:** IBM Netezza Analytics 2.x 이전 버전으로 모델이 작성된 경우 뷰어 탭이 비어 있습니다.

## IBM Data WH K-평균 모델 너깃

K-평균 모델 너깃은 훈련 데이터 및 추정 프로세스에 대한 정보와 함께, 군집 모델에서 캡처한 모든 정보를 포함합니다.

K-평균 모델 너깃을 포함하는 스트림을 실행하는 경우 노드는 해당 레코드의 지정된 군집 중심과의 거리 및 소속 군집을 포함하는 새 2개 필드를 추가합니다. 이름이 \$KM-K-Means인 새 필드는 소속군집용이며 이름이 \$KMD-K-Means인 새 필드는 군집 중심으로부터의 거리용입니다.

## IBM Data WH K-평균 너깃 - 모델 탭

모델 탭에는 군집 필드에 대한 분포 및 요약 통계를 표시하는 다양한 그래픽 보기가 있습니다. 모델에서 데이터를 내보내거나 보기를 그래픽으로 내보낼 수 있습니다.

IBM Netezza Analytics 2.x 이전 버전으로 작업 중이거나 거리 측도가 Mahalanobis의 거리인 모델을 작성하는 경우에는 의사결정 트리 모형의 내용이 텍스트 형식으로만 표시됩니다.

이러한 버전에 대해서는 다음 정보가 표시됩니다.

- 요약 통계량.** 가장 작은 군집 및 가장 큰 군집에 대해 요약 통계량이 레코드 수를 표시합니다. 또한 요약 통계량은 이러한 군집에 의해 사용된 데이터 세트의 퍼센트를 표시합니다. 또한 목록은 가장 큰 군집 대 가장 작은 군집의 크기 비율을 표시합니다.
- 군집 요약.** 군집 요약은 알고리즘에 의해 작성된 군집을 나열합니다. 테이블에는 각 군집에 대한 해당 군집 내의 레코드 수가 표시되며 해당 레코드에 대한 군집 중심으로부터의 평균 거리가 함께 표시됩니다.

## IBM Data WH K-평균 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**거리 측도.** 데이터 점 사이의 거리를 측정하는 데 사용할 방법입니다. 거리가 클수록 상이성도 커집니다. 옵션은 다음과 같습니다.

- 유클리디안.** (기본값) 두 점 사이의 거리는 두 점을 직선으로 결합하여 계산됩니다.
- 맨해튼.** 두 점 사이의 거리는 해당 좌표 간 절대 차이의 합계로 계산됩니다.
- 캔버라.** 맨해튼 거리와 비슷하지만 원점에 더 가까운 데이터 점에 더 민감합니다.
- 최대.** 두 점 사이의 거리가 좌표 차원을 따라 해당 차이 중 가장 큰 값으로 계산됩니다.

## Netezza Bayes Net 모델 너깃

Bayes Net 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

Bayes Net 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 14. Bayes Net에 대한 필드 모델 스코어링	
추가되는 필드의 이름	의미
\$BN-target_name	현재 레코드의 예측값입니다.

테이블 노드를 모델 너깃에 첨부하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

## Netezza Bayes Net 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**목표.** 현재 목표와 다른 대상 필드를 스코어링하려면 여기서 새 목표를 선택하십시오.

**레코드 ID.** 레코드 ID 필드가 지정되지 않은 경우 사용할 필드를 여기서 선택하십시오.

**예측 유형.** 사용할 예측 알고리즘의 변형입니다.

- 최적(상관관계가 가장 밀접한 이웃 항목).** (기본값) 상관관계가 가장 밀접한 이웃 항목 노드를 사용합니다.
- 이웃 항목(이웃 항목의 가중된 예측).** 모든 이웃 항목 노드의 가중된 예측을 사용합니다.
- NN-이웃 항목(null이 아닌 이웃 항목).** 널값인 노드(예측을 계산하는 대상이 되는 인스턴스에 대한 값이 누락된 속성에 해당하는 노드)를 무시하는 점만 제외하면 이전 옵션과 동일합니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

## IBM Data WH Naive Bayes 모델 너깃

Naive Bayes 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

Naive Bayes 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 15. Naive Bayes에 대한 모델 스코어링 필드 - 기본값	
추가되는 필드의 이름	의미
<code>\$I-target_name</code>	현재 레코드의 예측값입니다.

모델링 노드 또는 모델 너깃에서 스코어 레코드에 할당된 클래스의 확률 계산 옵션을 선택하고 스트림을 실행한 경우 두 개의 추가적 필드가 추가됩니다.

표 16. Naive Bayes에 대한 모델 스코어링 필드 - 추가적	
추가되는 필드의 이름	의미
<code>\$IP-target_name</code>	인스턴스에 대한 계층의 베이지안 문자입니다. 즉 사전 계층 확률 및 조건부 인스턴스 속성 값 확률의 곱입니다.
<code>\$ILP-target_name</code>	후자의 자연로그입니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

## IBM Data WH Naive Bayes 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**스코어링 레코드에 대해 지정된 클래스의 확률 계산.** (의사결정 트리 및 Naive Bayes 전용) 선택된 경우 이 옵션은 추가 모델링 필드에 신뢰도(즉, 확률) 필드 및 예측 필드가 포함됨을 의미합니다. 이 선택란을 선택 취소하면 예측 필드만 생성됩니다.

**적거나 심한 비균형 데이터 세트의 확률 정확성을 향상시킵니다.** 확률을 계산할 때 이 옵션이 추정 중에 0값 확률을 피하기 위한  $m$ -추정 기법을 사용합니다. 이런 종류의 확률 추정은 느릴 수는 있으나 적거나 심한 비균형 데이터 세트에 대해 더 나은 결과를 제공합니다.

## IBM Data WH KNN 모델 너깃

KNN 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

KNN 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 17. KNN에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
<code>\$KNN-target_name</code>	현재 레코드의 예측값입니다.

테이블 노드를 모델 너깃에 첨부하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

## IBM Data WH KNN 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**거리 측도.** 데이터 점 사이의 거리를 측정하는 데 사용할 방법입니다. 거리가 클수록 상이성도 커집니다. 옵션은 다음과 같습니다.

- **유클리디안.** (기본값) 두 점 사이의 거리는 두 점을 직선으로 결합하여 계산됩니다.
- **맨해튼.** 두 점 사이의 거리는 해당 좌표 간 절대 차이의 합계로 계산됩니다.
- **캔버라.** 맨해튼 거리와 비슷하지만 원점에 더 가까운 데이터 점에 더 민감합니다.
- **최대.** 두 점 사이의 거리가 좌표 차원을 따라 해당 차이 중 가장 큰 값으로 계산됩니다.

**최근접 이웃 수(k).** 특정 케이스에 대한 최근접 이웃 수입니다. 많은 수의 이웃을 사용한다고 해서 반드시 더 정확한 모델을 얻을 수 있는 것은 아님에 유의하십시오.

*k*를 선택하면 과적합 방지("불량" 데이터의 경우 특히 중요할 수 있음)와 해결(비슷한 인스턴스에 대해 다양한 예측을 생성함) 사이의 균형이 제어됩니다. 일반적으로 1부터 수십까지의 일반적인 값 범위를 사용하여 각 데이터 세트에 대해 *k*의 값을 조정해야 합니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**거리를 계산하기 전에 측정 표준화.** 선택된 경우 이 옵션은 거리 값을 계산하기 전에 연속 입력 필드에 대한 측정을 표준화합니다.

**큰 데이터 세트의 성능 향상을 위해 코어 세트 사용.** 선택된 경우 이 옵션은 코어 세트 표본 추출을 사용하여 큰 데이터 세트가 관련된 경우 계산 속도를 높입니다.

## Netezza 분열 군집 모델 너깃

분열 군집 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

분열 군집 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 두 개의 새 필드를 추가합니다.

표 18. 분할 군집에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
\$DC-target_name	현재 레코드가 지정되는 부군집의 식별자입니다.
\$DCD-target_name	현재 레코드에 대한 부군집 중심으로부터의 거리입니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

## Netezza 분열 군집 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**거리 측도.** 데이터 점 사이의 거리를 측정하는 데 사용할 방법입니다. 거리가 클수록 상이성도 커집니다. 옵션은 다음과 같습니다.

- **유클리디안.** (기본값) 두 점 사이의 거리는 두 점을 직선으로 결합하여 계산됩니다.
- **맨해튼.** 두 점 사이의 거리는 해당 좌표 간 절대 차이의 합계로 계산됩니다.
- **캔버라.** 맨해튼 거리와 비슷하지만 원점에 더 가까운 데이터 점에 더 민감합니다.
- **최대.** 두 점 사이의 거리가 좌표 차원을 따라 해당 차이 중 가장 큰 값으로 계산됩니다.

적용되는 계층 수준. 데이터에 적용되어야 하는 계층 구조의 수준입니다.

## IBM Data WH PCA 모델 너깃

PCA 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

PCA 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 19. PCA에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
<code>\$F-target_name</code>	현재 레코드의 예측값입니다.

모델링 노드 또는 모델 너깃의 주성분 수... 필드에서 1보다 큰 값을 지정하고 스트림을 실행하면 노드가 각 성분에 대해 새 필드를 추가합니다. 이 경우, 필드 이름에  $n$  접미문자가 추가되며  $n$ 은 성분의 번호입니다. 예를 들어, 모델 이름이 `pca`이며 세 개의 성분이 있는 경우, 새 필드 이름이 `$F-pca-1`, `$F-pca-2` 및 `$F-pca-3`이 됩니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

**참고:** 소문자 필드 이름을 사용할 경우 모델을 스코어링할 때 오류가 발생할 수 있습니다. 이것은 알려진 Db2 Data Warehouse 결함으로, 임시 해결책은 스코어링하기 전에 모든 필드의 이름을 대문자로 변경하는 것입니다.

## IBM Data WH PCA 너깃 - 설정 탭

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**투영에 사용될 주성분의 수.** 데이터 세트를 줄일 주성분의 수입니다. 이 값은 속성의 수(입력 필드)를 초과할 수 없습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

## Netezza 회귀분석 트리 모델 너깃

회귀분석 트리 모델 너깃은 모델 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

회귀분석 트리 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 20. 회귀분석 트리에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
<code>\$I-target_name</code>	현재 레코드의 예측값입니다.

모델링 노드 또는 모델 너깃에서 **추정 분산 계산** 옵션을 선택하고 스트림을 실행한 경우 추가적 필드가 추가됩니다.

표 21. 회귀분석 트리에 대한 모델 스코어링 필드 - 추가적	
추가되는 필드의 이름	의미
<code>\$IV-target_name</code>	예측값의 추정 분산입니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

## Netezza 회귀분석 트리 너깃 - 모델 탭

모델 탭은 회귀분석 트리 모델의 예측변수 중요도를 그래픽 형식으로 표시합니다. 막대의 길이는 예측변수의 중요도를 나타냅니다.

**참고:** IBM Netezza Analytics 2.x 이전 버전으로 작업 중인 경우에는 회귀분석 트리 모델의 내용이 텍스트 형식으로만 표시됩니다.

이러한 버전에 대해서는 다음 정보가 표시됩니다.

- 노드 또는 리프에 해당하는 텍스트의 각 줄
- 트리 수준을 반영하는 들여쓰기
- 노드의 경우, 분할 조건이 표시됩니다.
- 리프의 경우, 지정된 클래스 레이블이 표시됩니다.

### **Netezza 회귀분석 트리 너깃 - 설정 탭**

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

**추정 분산 계산.** 지정된 클래스의 분산이 출력에 포함되어야 하는지 여부를 표시합니다.

### **Netezza 회귀분석 트리 너깃 - 뷰어 탭**

뷰어 탭은 SPSS Modeler가 해당 회귀분석 트리 모델의 트리 프리젠테이션을 표시하는 것과 동일한 방식으로 트리 모델의 트리 프리젠테이션을 표시합니다.

**참고:** IBM Netezza Analytics 2.x 이전 버전으로 모델이 작성된 경우 뷰어 탭이 비어 있습니다.

## **IBM Data WH 선형 회귀 모델 너깃**

선형 회귀 모형 너깃은 모형 스코어링에 필요한 옵션을 설정하는 방법을 제공합니다.

선형 회귀 모형 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 한 개의 새 필드를 추가합니다.

표 22. 선형 회귀에 대한 모델 스코어링 필드

추가되는 필드의 이름	의미
\$LR-target_name	현재 레코드의 예측값입니다.

### **IBM Data WH 선형 회귀 너깃 - 설정 탭**

설정 탭에서 모델 스코어링에 필요한 옵션을 설정할 수 있습니다.

**입력 필드 포함.** 선택된 경우 이 옵션은 모든 원래 입력 필드 다운스트림을 전달하여 추가 모델링 필드를 각 데이터 행에 추가합니다. 이 선택란을 선택 취소하면 레코드 ID 필드 및 추가 모델링 필드만 전달되므로 스트림이 더 신속하게 실행됩니다.

## **Netezza 시계열 모델 너깃**

모델 너깃은 시계열 모델링 작업의 출력에 대한 액세스를 제공합니다. 출력은 다음 필드로 구성됩니다.

표 23. 시계열 모델 출력 필드

필드	설명
TSID	시계열의 식별자이며 모델링 노드의 필드 탭의 시계열 ID에 대해 지정되는 내용입니다. 자세한 정보는 60 페이지의 『Netezza 시계열 필드 옵션』의 내용을 참조하십시오.
시간	현재 시계열 내의 시간 주기입니다.

표 23. 시계열 모델 출력 필드 (계속)	
필드	설명
히스토리	히스토리 데이터 값(예측 작성에 사용된 값)입니다. 이 필드는 모델 너깃의 설정 탭에서 결과에 히스토리 값 포함 옵션이 선택된 경우에만 포함됩니다.
\$TS-INTERPOLATED	보간법이 적용된 값이며 사용된 경우에 한합니다. 이 필드는 모델 너깃의 설정 탭에서 결과에 보간값 포함 옵션이 선택된 경우에만 포함됩니다. 보간법은 모델링 노드의 작성 옵션 탭의 옵션입니다.
\$TS-FORECAST	시계열의 예측값입니다.

모델 출력을 보려면 테이블 노드를 노드 팔레트의 출력 탭에서 모델 너깃으로 연결하고 테이블 노드를 실행하십시오.

### Netezza 시계열 너깃 - 설정 탭

설정 탭에서 모델 출력 사용자 정의에 필요한 옵션을 설정할 수 있습니다.

**모델 이름.** 모델링 노드의 모델 옵션 탭에서 지정된 모델의 이름입니다.

기타 옵션은 모델링 노드의 모델링 옵션 탭과 동일합니다.

## IBM Data WH 일반화 선형 모델 너깃

모델 너깃은 모델링 작업의 출력에 대한 액세스를 제공합니다.

일반화 선형 모형 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 새 필드를 추가합니다.

표 24. 일반화 선형에 대한 모델 스코어링 필드

추가되는 필드의 이름	의미
\$GLM-target_name	현재 레코드의 예측값입니다.

모델 탭에는 모델과 연관된 다양한 통계량이 표시됩니다.

출력은 다음 필드로 구성됩니다.

표 25. 일반화 선형 모형의 출력 필드

출력 필드	설명
매개변수	모델에서 사용되는 모수(예측자 변수)입니다. 절편(회귀 모형의 상수항)과 같이 수치 및 명목형 열입니다.
베타	상관계수(모델의 선형 성분)입니다.
표준오차	베타에 대한 표준 편차입니다.
검정	모수의 타당성을 평가하는 데 사용되는 검정 통계량입니다.
P-값	모수가 유의적이라고 가정할 때 오차의 확률입니다.
<b>잔차 요약</b>	
잔차 유형	요약 값이 표시되는 예측의 잔차 유형입니다.
RSS	잔차 값입니다.
자유도	잔차에 대한 자유도입니다.
P-값	오차의 확률입니다. 높은 값은 적합도가 낮은 모델을 나타내며 낮은 값은 적합도가 높은 모델을 나타냅니다.

## **IBM Data WH 일반화 선형 모형 너깃 - 설정 탭**

설정 탭에서 모델 출력을 사용자 정의할 수 있습니다.

이 옵션은 모델링 노드의 스코어링 옵션에 대해 표시된 것과 동일합니다. 자세한 정보는 51 페이지의 [『IBM Data WH 일반화 선형 모델 옵션 - 스코어링 옵션』](#)의 내용을 참조하십시오.

## **IBM Data WH TwoStep 모델 너깃**

TwoStep 모델 너깃을 포함하는 스트림을 실행하는 경우, 노드는 소속군집 및 해당 레코드에 대해 지정된 군집 중심으로부터의 거리를 포함하는 두 개의 새 필드를 추가합니다. 이름이 \$TS-Twostep인 새 필드는 소속군집용이며 이름이 \$TSP-Twostep인 새 필드는 군집 중심으로부터의 거리용입니다.

## **IBM Data WH TwoStep 너깃 - 모델 탭**

모델 탭에는 군집 필드에 대한 분포 및 요약 통계를 표시하는 다양한 그래픽 보기와 모델에서 데이터를 내보내거나 보기와 같은 기능이 있습니다.

# 제 6 장 z/OS용 IBM Db2를 사용한 데이터베이스 모델링

## IBM SPSS Modeler 및 z/OS용 IBM Db2

SPSS Modeler는 z/OS용 Db2와의 통합을 지원하여 z/OS용 Db2 서버에서 고급 분석을 실행할 수 있습니다. SPSS Modeler 그래픽 사용자 인터페이스 및 워크플로우 지향 개발 환경을 통해 이러한 기능에 액세스할 수 있습니다. 이 방식으로, IBM Db2 Analytics Accelerator를 활용하여 z/OS용 Db2 환경에서 직접 데이터 마이닝 알고리즘을 실행할 수 있습니다.

SPSS Modeler는 z/OS용 Db2로부터 다음과 같은 알고리즘의 통합을 지원합니다.

- 의사결정 트리
- K-평균
- Naive Bayes
- 회귀분석 트리
- TwoStep

## z/OS용 IBM Db2와의 통합을 위한 요구사항

다음 조건은 z/OS®용 Db2® 및 z/OS용 IBM Db2 Analytics Accelerator를 사용하여 In-Database 모델링을 수행하기 위한 전제조건입니다. 이러한 조건이 충족되었는지 확인하기 위해 데이터베이스 관리자와 상의할 필요가 있습니다. 지원되는 버전을 포함하는 자세한 요구사항은 [소프트웨어 제품 호환성 보고서](#)를 참조하십시오.

- Windows 또는 UNIX에서 SPSS Modeler Server 설치에 대해 또는 로컬 모드로 실행 중인 IBM SPSS Modeler
- z/OS용 Db2 Analytics Accelerator가 있는 z/OS용 Db2
- IBM SPSS Data Access Pack
- SPSS Modeler Server를 실행 중인 서버에서, 다음 시스템 중 하나:
  - IBM Db2 Data Server Driver for ODBC and CLI
  - Linux®, UNIX 및 Windows용 Db2 버전과 z/OS용 Db2에 맞게 구성된 ODBC 데이터 소스
- Db2 Connect for System z®에 대한 라이센스
- SPSS Modeler에서 SQL 생성 및 최적화 사용
- Db2 z/OS In-Database 마이닝에는 액셀러레이터 전용 테이블(AOT) 또는 가속화된 테이블 및 INZA 지원이 필요합니다. IDAA INZA는 IDAA 5.1에서 도입되었습니다. 즉, Db2 z/OS In-Database 마이닝 노드는 이전 버전의 IDAA에서는 작동하지 않습니다.

Modeler에서 IDAA 사용 가능 DSN을 사용하는 경우, 해당 DSN을 사용하는 데이터베이스 소스 노드에서 리턴되는 테이블의 목록에 표시될 유일한 테이블은 AOT 또는 가속화된 테이블입니다.

## z/OS용 IBM Db2 Analytics Accelerator와의 통합 사용

z/OS용 Db2 Analytics Accelerator와의 통합을 사용하기 위한 단계는 다음과 같이 구성됩니다.

- z/OS용 Db2 및 z/OS용 Db2 Analytics Accelerator 구성
- ODBC 소스 작성
- IBM SPSS Modeler에서 z/OS용 IBM Db2의 통합 사용
- SPSS Modeler에서 SQL 생성 및 최적화 사용
- z/OS용 Db2에 대해 IBM SPSS Modeler Server Scoring Adapter 사용
- IBM SPSS Modeler에서 IBM Db2 클라이언트를 사용하여 DSN 구성

## **z/OS용 IBM Db2 및 z/OS용 IBM Analytics Accelerator 구성**

z/OS용 Db2 및 z/OS용 Analytics Accelerator를 구성하는 방법은 다음 웹 사이트에 설명되어 있습니다.

[Db2 Analytics Accelerator for z/OS](#).

## **z/OS용 IBM Db2 및 IBM Db2 Analytics Accelerator에 대한 ODBC 소스 작성**

z/OS용 Db2와 IBM Db2 Analytics Accelerator 사이의 연결 활성화 방법에 대한 정보는 다음 웹 사이트를 참조하십시오.

- 버전 4의 경우: [Db2 Analytics Accelerator for z/OS 4.1.0](#)
- 버전 3의 경우: [Db2 Analytics Accelerator for z/OS 3.1.0](#)
- 애플리케이션에 대한 수정 없이 ODBC 및 JDBC 애플리케이션에서 IBM Db2 Analytics Accelerator로 쿼리 가속화 사용
- z/OS용 Db2 Analytics Accelerator에서 쿼리를 실행할 때 ODBC 드라이버의 SQL 오류

## **IBM SPSS Modeler에서 z/OS용 IBM Db2의 통합 사용**

SPSS Modeler에서 z/OS용 Db2의 통합을 사용하려면 다음 단계를 수행하십시오.

1. SPSS Modeler config 디렉토리에서 odbc-db2-accelerator-names.cfg 파일을 여십시오.

파일이 존재하지 않으면 파일을 작성해야 합니다.

2. 모든 데이터 소스 이름과 모든 액셀러레이터 이름을 추가하십시오. 예를 들어,

```
dsn1, acceleratorname1  
dsn2, acceleratorname2
```

3. 액셀러레이터 전용 테이블(AOT)의 기본 CCSID는 유니코드입니다. 이를 대체하려면 액셀러레이터 이름에 인코딩 문자열을 추가하여 항목을 수정하십시오. 예를 들어,

```
dsn1, acceleratorname1, EBCDIC  
dsn2, acceleratorname2, UNICODE
```

4. odbc-db2-accelerator-names.cfg 파일을 저장하고 닫은 후, 동일한 디렉토리에서 odbc-db2-custom-properties.cfg 파일을 여십시오.

5. SPSS Modeler는 SQL을 사용하여 IDAA 레지스터를 설정합니다. 필요한 경우에는 SQL을 필요한 값으로 변경하여 이러한 항목을 대체할 수 있습니다. 예를 들어,

```
current_query_sql_acc, "SET CURRENT QUERY ACCELERATION = ELIGIBLE"  
current_get_archive_acc, "SET CURRENT GET_ACCEL_ARCHIVE = NO"
```

6. 기본적으로 SPSS Modeler는 SQL을 사용하여 데이터베이스 캐시용 임시 테이블을 작성합니다. 필요한 경우에는 원하는 데이터베이스 이름을 지정하여 이를 대체할 수 있습니다. 예를 들어,

```
[OSZ]  
table_create_temp_sql_acc, 'CREATE TABLE <table-name> <(table-columns)> IN DATABASE  
NAME_OF_DATABASE_FOR_AOT'
```

7. 기본적으로 SPSS Modeler는 ODBC 소스 노드에 작성된 SQL 쿼리를 재실행 불가능으로 간주하며, 이는 해당 쿼리가 여러 번 실행되는 경우 서로 다른 결과를 리턴할 것으로 예상됨을 의미합니다. 그러나 일부 시나리오에서는 이로 인해 Modeler가 다운스트림 노드에 대해 SQL을 생성하지 못할 수 있으며 이는 관련 값을 Y로 변경하여 대체할 수 있습니다. 예:

```
assume_custom_sql_replayable, Y
```

8. SPSS Modeler 주메뉴에서 도구 > 옵션 > 헬퍼 애플리케이션을 클릭하십시오.

9. **z/OS용 IBM Db2** 탭을 클릭하십시오.

10. **z/OS용 IBM Db2 데이터 마이닝 통합 사용**을 선택한 후 확인을 클릭하십시오.

참고: Modeler에서 IDAA 테이블과 비IDAA 테이블을 동시에 볼 수 없습니다.

## SQL 생성 및 최적화 사용

매우 큰 데이터 세트에 대해 작업할 수도 있기 때문에 성능을 위해 IBM SPSS Modeler에서 SQL 생성 및 최적화 옵션을 사용으로 설정해야 합니다.

SPSS Modeler를 구성하려면 다음 단계를 수행하십시오.

1. IBM SPSS Modeler 메뉴에서 도구 > 스트림 특성 > 옵션을 선택하십시오.
2. 탐색 분할창에서 최적화 옵션을 클릭하십시오.
3. **SQL 생성** 옵션이 사용으로 설정되어 있는지 확인하십시오. 이 설정은 데이터베이스 모델링이 작동하기 위해 필수입니다.
4. **SQL 생성 최적화 및 기타 실행 최적화**를 선택하십시오(엄격하게 요구되지 않지만 최적화된 성능을 위해서는 강력하게 권장됨).

## IBM SPSS Modeler에서 IBM Db2 클라이언트를 사용하여 DSN 구성

필요한 경우, SPSS Modeler에서 Db2용 Db2 클라이언트를 사용하여 데이터 소스 이름(DSN)을 구성하려면 다음 단계를 완료하십시오.

1. 아직 설치되지 않은 경우, Modeler Server가 설치된 운영 체제에 Db2 클라이언트를 설치하십시오.
2. **db2 catalog** 명령을 사용하여 데이터베이스를 카탈로그화하고 새 데이터 소스를 Db2 클라이언트 내의 db2cli.ini 파일에 추가하십시오. 정의된 데이터베이스 별명을 지정하는지 확인하십시오.
3. 데이터 액세스를 구성하십시오. 자세한 단계는 Modeler 문서에 있습니다.
4. 2단계에서 정의된 데이터베이스 별명을 참조하여 odbc.ini에 새 ODBC 데이터 소스를 작성하십시오.
5. Linux 또는 UNIX 사용자의 경우:
  - a. 드라이버 라이브러리 libdb2o.so가 libdb2.so 대신 사용되었으며 새 데이터 소스에 대해 'DriverUnicodeType=1'이 정의되었는지 확인하십시오.
  - b. IBM SPSS Data Access Pack 설치에서 Db2 클라이언트의 라이브러리 경로가 odbc.sh에 추가되었는지 확인하십시오.
  - c. Modeler Server가 UTF-16 인코딩을 사용하는 ODBC 드라이버 랩퍼 라이브러리 ('libspssodbc\_datadirect\_utf16.so')를 사용하는지 확인하십시오.
6. Db2에 연결하는 사용자가 다음 쿼리를 실행할 수 있는 권한을 갖고 있는지 확인하십시오.

```
SELECT ACCELERATORNAME FROM SYSACCEL.SYSACCELERATORS
```

## z/OS용 IBM Db2를 사용한 모델 작성

각각의 지원되는 알고리즘에는 해당 모델링 노드가 있습니다. 노드 팔레트의 데이터베이스 모델링 탭에서 z/OS 용 Db2 모델링 노드에 액세스할 수 있습니다.

### 데이터 고려사항

데이터 소스에 있는 필드는 모델링 노드에 따라 다양한 데이터 유형의 변수를 포함할 수 있습니다. SPSS Modeler에서는 데이터 유형이 측정 수준으로 알려져 있습니다. 모델링 노드의 필드 탭에서는 아이콘을 사용하여 해당 입력 및 대상 필드에 대해 허용되는 측정 수준 유형을 표시합니다.

**대상 필드.** 대상 필드는 값을 예측하는 필드입니다. 목표를 지정할 수 있는 경우 소스 데이터 필드 중 하나만 대상 필드로 선택할 수 있습니다.

**레코드 ID 필드.** 각각의 케이스를 고유하게 식별하는 데 사용되는 필드를 지정합니다. 예를 들어, ID 필드(예: CustomerID)가 될 수 있습니다. 소스 데이터가 ID 필드를 포함하지 않는 경우에는 다음 프로시저에 표시된 대로 파생 노드를 사용하여 이 필드를 작성할 수 있습니다.

1. 소스 노드를 선택하십시오.
2. 노드 팔레트의 필드 조작 탭에서 파생 노드를 두 번 클릭하십시오.
3. 캔버스에서 해당 아이콘을 두 번 클릭하여 파생 노드를 여십시오.
4. 파생 필드 필드에 예를 들어, ID를 입력하십시오.
5. 수식 필드에서 @INDEX를 입력한 후 확인을 클릭하십시오.
6. 파생 노드를 나머지 스트림에 연결하십시오.

## 널값 처리

입력 데이터에 널값이 포함되어 있는 경우 일부 z/OS용 Db2 노드를 사용하면 오류 메시지가 표시되거나 장기 실행 스트림이 발생할 수 있으므로 널값이 포함된 레코드는 제거하는 것이 좋습니다. 다음의 방법을 사용하십시오.

1. 선택 노드를 소스 노드에 연결하십시오.
2. 선택 노드의 모드 옵션을 삭제로 설정하십시오.
3. 조건 필드에서 다음을 입력하십시오.

```
@NULL(field1) [or @NULL(field2)[... or @NULL(fieldN)]]
```

- 모든 입력 필드를 포함해야 합니다.
4. 선택 노드를 나머지 스트림에 연결하십시오.

## 모델 출력

z/OS용 Db2 모델링 노드가 포함된 스트림이 실행될 때마다 약간 다른 결과를 생성할 수 있습니다. 이는 모델 작성 전에 데이터를 임시 테이블로 읽어오므로 노드가 소스 데이터를 읽는 순서가 항상 동일하지 않기 때문입니다. 하지만 이 영향에 의해 생성된 차이는 무시할 수 있습니다.

## 일반 주석

- SPSS Collaboration and Deployment Services에서는 z/OS용 Db2 모델링 노드가 포함된 스트림을 사용하여 스코어링 구성을 작성할 수 없습니다.
- z/OS용 Db2 노드에 의해 작성된 모델에 대해서는 PMML 내보내기 또는 가져오기를 수행할 수 없습니다.

## z/OS용 IBM Db2 모델 - 필드 옵션

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.**

**목표.** 예측에 대한 목표로 하나의 필드를 선택하십시오. 일반화 선형 모형의 경우 이 화면의 시행 필드도 참조하십시오.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## **z/OS용 IBM Db2 모델 - 서버 옵션**

서버 탭에서 모델을 작성할 z/OS용 Db2 시스템을 지정합니다.

- 업스트림 연결 사용.** (기본값) 업스트림 노드(예: 데이터베이스 소스 노드)에 지정된 연결 세부사항을 사용합니다. 참고: 이 옵션은 모든 업스트림 노드가 SQL 푸시백을 사용할 수 있는 경우에만 작동합니다. 이 경우에는 SQL이 모든 업스트림 노드를 완전하게 구현하므로 데이터를 데이터베이스 밖으로 이동하지 않아도 됩니다.
- 데이터를 연결로 이동.** 여기에서 지정하는 데이터베이스로 데이터를 이동합니다. 이를 수행하면 데이터가 다른 IBM 데이터베이스 또는 다른 벤더의 데이터베이스에 있거나 데이터가 플랫 파일인 경우에도 모델링이 작동할 수 있습니다. 또한, 노드가 SQL 푸시백을 수행하지 않아 데이터가 추출된 경우에는 여기에 지정된 데이터베이스로 데이터가 다시 이동합니다. 편집 단추를 클릭하여 연결을 찾아서 선택할 수 있습니다.

**참고:** ODBC 데이터 소스 이름이 각 SPSS Modeler 스트림에 효과적으로 임베드됩니다. 한 호스트에서 작성된 스트림이 다른 호스트에서 실행되는 경우 데이터 소스의 이름이 각 호스트에서 동일해야 합니다. 또는 각 소스 또는 모델링 노드의 서버 탭에서 다른 데이터 소스를 선택할 수 있습니다.

## **z/OS용 IBM Db2 모델 - 모델 옵션**

모델 옵션 탭에서는 모델 이름을 지정할 것인지, 이름을 자동으로 생성할 것인지 선택할 수 있습니다.

**모델 이름.** 목표나 ID 필드(또는 이러한 필드가 지정되지 않은 경우에는 모델 유형)를 기준으로 하여 모델 이름을 자동으로 생성하거나 사용자 정의 이름을 지정할 수 있습니다.

**이름이 사용된 경우 기존 이름 바꾸기.** 이 선택란을 선택하면 동일한 이름을 가진 모든 기존 모델을 덮어씁니다.

## **z/OS용 IBM Db2 모델 - K-평균**

K-평균 노드는 군집분석 방법을 제공하는 K-평균 알고리즘을 구현합니다. 이 노드를 사용하여 데이터 세트를 별개의 그룹으로 군집화할 수 있습니다.

이 알고리즘은 거리 메트릭(함수)을 기반으로 데이터 점 사이의 유사성을 측정하는 거리 기반 군집화 알고리즘입니다. 데이터 점은 사용되는 거리 메트릭에 따라 가장 가까운 군집에 지정됩니다.

각 학습 인스턴스가 가장 가까운 군집에 지정되는 동일한 기본 프로세스를 여러 번 반복 수행함으로써 이 알고리즘이 작동합니다(지정된 거리 함수와 관련하여 해당 인스턴스 및 군집 중심에 적용됨). 그러면 모든 군집 중심이 특정 군집에 지정된 인스턴스의 평균 속성 값 벡터로서 다시 계산됩니다.

## **z/OS용 IBM Db2 모델 - K-평균 필드 옵션**

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.**

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## **z/OS용 IBM Db2 모델 - K-평균 작성 옵션**

작성 옵션을 설정하여 고유의 목적에 맞게 모델 작성을 사용자 정의할 수 있습니다.

**기본 옵션을 사용하여 모델을 작성하려면 실행을 클릭하십시오.**

**거리 측도.** 이 매개변수는 데이터 점 사이의 거리를 측정하는 방법을 정의합니다. 거리가 멀수록 유사성이 떨어집니다. 다음 옵션 중 하나를 선택하십시오.

- **유클리디안.** 유클리디안 측도는 두 데이터 점 사이의 직선 거리입니다.
- **정규화 유클리디안.** 정규화 유클리디안 측도는 유클리디안 측도와 유사하지만 제곱 표준 편차에 의해 정규화 됩니다. 유클리디안 측도와 달리, 정규화 유클리디안 측도는 척도 불변성(scale-invariant)을 가집니다.

**군집 수.** 이 매개변수는 작성할 군집 수를 정의합니다.

**반복 최대 수.** 알고리즘이 동일한 프로세스를 여러 번 반복합니다. 이 매개변수는 모델 학습이 중지하는 반복 수를 정의합니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.

- **없음.** 모델을 스코어링하는데 필요한 통계량만 포함됩니다.

**결과 복제.** 분석을 복제하기 위한 난수 시드를 설정하려면 이 선택란을 선택하십시오. 정수를 지정하거나 **생성을 클릭하여** 의사 난수 정수를 작성할 수 있습니다.

## **z/OS용 IBM Db2 모델 - Naive Bayes**

Naive Bayes는 분류 문제점을 위한 잘 알려진 알고리즘입니다. 모델은 제안된 모든 예측 변수를 서로 독립된 것으로 처리하기 때문에 naïve라고 합니다. Naive Bayes는 속성과 목표 속성의 조합에 대한 조건부 확률을 계산하는 빠르고 확장 가능한 알고리즘입니다. 학습 데이터로부터 독립적인 확률이 설정됩니다. 이 확률은 각 입력 변수에서 각각의 값 범주가 발생하는 경우 각 목표 클래스의 우도를 제공합니다.

## **z/OS용 IBM Db2 모델 - 의사결정 트리**

의사결정 트리는 분류 모델을 나타내는 계층 구조입니다. 의사결정 트리 모형을 사용하면 학습 데이터 세트에서 미래 관측을 예측하거나 분류하는 분류 시스템을 개발할 수 있습니다. 분류는 분류의 분할 포인트를 표시하는 가지가 있는 트리 구조 형식을 사용합니다. 분할은 중지 포인트에 도달할 때까지 반복적으로 데이터를 하위 그룹으로 분류합니다. 중지 포인트의 트리 노드를 리프라고 합니다. 각 리프는 하위 그룹 또는 클래스의 멤버에 클래스 레이블로 알려진 레이블을 지정합니다.

## **z/OS용 IBM Db2 모델 - 의사결정 트리 필드 옵션**

필드 탭에서는 업스트림 노드에 이미 정의된 필드 역할 설정을 사용하거나 수동으로 필드 할당을 할 수 있습니다.

**사전 정의된 역할 사용.** 이 옵션은 업스트림 유형 노드(또는 업스트림 소스 노드의 유형 탭)의 역할 설정(목표, 예측변수 등)을 사용합니다.

**사용자 정의 필드 할당 사용.** 이 화면에서 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표 단추를 사용하여 수동으로 이 목록의 항목을 화면 오른쪽의 다양한 역할 필드에 할당하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**모두 단추를 클릭하여** 목록의 모든 필드를 선택하거나 개별 측정 수준 단추를 클릭하여 해당 측정 수준의 모든 필드를 선택합니다.

**목표.** 예측에 대한 목표로 하나의 필드를 선택하십시오.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다. 이 필드의 값은 레코드마다 고유해야 합니다(예: 고객 ID 번호).

**인스턴스 가중치.** 여기에 필드를 지정하면 기본값인 클래스 가중치(대상 필드에 대한 범주당 가중치) 대신 또는 이와 더불어 인스턴스 가중치(입력 데이터의 행당 가중치)를 사용할 수 있습니다. 여기서 지정하는 필드는 입력 데이터의 각 행에 대한 숫자 가중치를 포함하는 필드여야 합니다.

**예측변수(입력).** 하나 이상의 입력 필드를 선택합니다. 유형 노드에서 필드 역할을 입력으로 설정하는 것과 유사합니다.

## **z/OS용 IBM Db2 모델 - 의사결정 트리 작성 옵션**

다음 작성 옵션을 트리 성장에 사용할 수 있습니다.

**확장 측도.** 이 옵션은 트리 성장이 측정되는 방법을 제어합니다.

- **불순도 측도.** 이 측도는 트리를 분할하기에 가장 적합한 위치를 평가합니다. 데이터 하위 그룹 또는 세그먼트에 서의 변동 측정치입니다. 낮은 불순도 측정치는 대부분의 멤버가 기준 또는 대상 필드에 대해 유사한 값을 갖는 그룹을 나타냅니다.

지원되는 측정은 **엔트로피** 및 **지니(Gini)**입니다. 이러한 측정은 분기에 대한 범주 소속 확률을 기반으로 합니다.

- **최대 나무 깊이.** 루트 노드 아래에서 트리가 확장될 수 있는 최대 수준 수, 즉 표본이 반복적으로 분할되는 횟수입니다. 이 특성의 기본값은 10이며 이 특성에 대해 설정할 수 있는 최대값은 62입니다.

**참고:** 모델 너깃의 뷰어가 모델을 텍스트 형식으로 표현하는 경우 최대 12수준의 트리가 표시됩니다.

**분할 기준.** 이 옵션은 트리 분할 중지 시점을 제어합니다.

- **분할할 최소 개선도.** 트리에서 새 분할이 작성되기 전에 제거되어야 하는 불순도의 최소 양입니다. 트리 작성의 목표는 유사한 출력 값을 가진 하위 그룹을 작성하여 각 노드 내의 불순도를 최소화하는 것입니다. 분기에 대한 최상의 분할이 분할 기준에 의해 지정된 수치 미만으로 불순도를 감소시키는 경우 분기가 분할되지 않습니다.
- **분할할 인스턴스의 최소수.** 분할할 수 있는 최소 레코드 수입니다. 분할되지 않은 레코드가 이 수 미만으로 남아 있는 경우, 추가 분할이 수행되지 않습니다. 이 필드를 사용하여 트리에서 작은 하위 그룹이 작성되는 것을 방지할 수 있습니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.

- **없음.** 모델을 스코어링하는데 필요한 통계량만 포함됩니다.

## **z/OS용 IBM Db2 모델 - 의사결정 트리 노드 - 클래스 가중치**

여기서는 개별 클래스에 가중치를 지정할 수 있습니다. 기본값은 모든 클래스에 동일한 가중치가 부여되도록 모든 클래스에 1 값을 지정하는 것입니다. 서로 다른 클래스 레이블에 대해 서로 다른 숫자 가중치를 지정함으로써 알고리즘이 그에 따라 특정 클래스의 학습 세트에 가중치를 주도록 합니다.

가중치를 변경하려면 **가중치** 열에서 가중치를 두 번 클릭하고 원하는 대로 변경하십시오.

**값.** 대상 필드의 가능한 값에서 파생된 클래스 레이블 세트입니다.

**가중치.** 특정 클래스에 지정할 가중치입니다. 클래스에 지정하는 가중치가 높을수록 모델이 해당 클래스에 대해 다른 클래스보다 더 민감하게 반응합니다.

클래스 가중치와 인스턴스 가중치를 조합하여 사용할 수 있습니다.

## **z/OS용 IBM Db2 모델 - 의사결정 트리 노드 - 트리 가지치기**

가지치기 옵션을 사용하여 의사결정 트리의 가지치기 기준을 지정할 수 있습니다. 가지치기의 목적은 새 데이터에 대한 예상 정확도를 향상시키지 않는 지나치게 확장된 하위 그룹을 제거하여 과적합 위험을 줄이기 위한 것입니다.

**가지치기 측도.** 기본 가지치기 측도인 **정확도**는 트리에서 리프를 제거한 후 모델의 추정 정확도가 허용 가능한 한 계 내로 유지되도록 합니다. 가지치기를 적용하는 동안 클래스 가중치를 고려하려면 대안인 **가중 정확도**를 사용하십시오.

**가지치기를 위한 데이터.** 학습 데이터의 일부 또는 모두를 사용하여 새 데이터에 대한 예상 정확도를 추정할 수 있습니다. 또는 이 용도로 지정된 테이블에서 별도의 가지치기 데이터 세트를 사용할 수 있습니다.

- **모든 학습 데이터 사용.** 이 옵션(기본값)은 모든 학습 데이터를 사용하여 모형 정확도를 추정합니다.
- **가지치기를 위해 학습 데이터의 % 사용.** 가지치기 데이터에 대해 여기서 지정된 백분율을 사용하여 데이터를 두 개의 세트(학습에 대한 세트 하나와 가지치기에 대한 세트 하나)로 분할하려면 이 옵션을 사용하십시오.
- 난수 시드를 지정하여 스트림을 실행할 때마다 데이터가 동일한 방식으로 파티셔닝되도록 하려면 **결과 복제**를 선택하십시오. 가지치기에 사용되는 시드 필드에 정수를 지정하거나 **생성**을 클릭하여 의사 난수 정수를 작성할 수 있습니다.
- **기존 테이블의 데이터 사용.** 모형 정확도 추정을 위한 별도의 가지치기 데이터 세트의 테이블 이름을 지정하십시오. 이 방법은 학습 데이터를 사용하는 것보다 신뢰성이 높은 것으로 간주됩니다.

## **z/OS용 IBM Db2 모델 - 회귀분석 트리**

회귀분석 트리는 케이스 표본을 반복적으로 분할하여 숫자 대상 필드의 값을 기반으로 동일한 종류의 서브세트를 파생하는 트리 기반 알고리즘입니다. 의사결정 트리와 마찬가지로, 회귀분석 트리는 트리의 리프가 충분히 작거나 충분히 균일한 서브세트에 해당되는 여러 서브세트로 데이터를 분해합니다. 분할은 목표 속성 값의 산포도를 줄이기 위해 선택합니다. 그러면 리프에서의 해당 평균 값을 사용하여 목표 속성 값을 상당히 잘 예측할 수 있습니다.

## **z/OS용 IBM Db2 모델 - 회귀분석 트리 작성 옵션 - 트리 성장**

트리 성장 및 트리 가지치기를 위한 작성 옵션을 설정할 수 있습니다.

다음 작성 옵션을 트리 성장에 사용할 수 있습니다.

**최대 나무 깊이.** 루트 노드 아래에서 트리가 확장될 수 있는 최대 수준 수, 즉 표본이 반복적으로 분할되는 횟수입니다. 기본값은 62이며, 이 값은 모델링을 위한 최대 트리 깊이입니다.

**참고:** 모델 너깃의 뷰어가 모델을 텍스트 형식으로 표현하는 경우 최대 12수준의 트리가 표시됩니다.

**분할 기준.** 이 옵션은 트리 분할 중지 시점을 제어합니다.

- **분할 평가 측도.** 이 클래스 평가 측도는 트리를 분할하기에 가장 적합한 위치를 평가합니다.

**참고:** 현재, 분산이 사용 가능한 유일한 옵션입니다.

- **분할할 최소 개선도.** 트리에서 새 분할이 작성되기 전에 제거되어야 하는 불순도의 최소 양입니다. 트리 작성의 목표는 유사한 출력 값을 가진 하위 그룹을 작성하여 각 노드 내의 불순도를 최소화하는 것입니다. 분기에 대한 최상의 분할이 분할 기준에 의해 지정된 수치 미만으로 불순도를 감소시키는 경우 분기가 분할되지 않습니다.
- **분할할 인스턴스의 최소수.** 분할할 수 있는 최소 레코드 수입니다. 분할되지 않은 레코드가 이 수 미만으로 남아 있는 경우, 추가 분할이 수행되지 않습니다. 이 필드를 사용하여 트리에서 작은 하위 그룹이 작성되는 것을 방지할 수 있습니다.

**통계량.** 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 다음 옵션 중 하나를 선택하십시오.

- **모두.** 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

**참고:** 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 **없음**을 지정하십시오.

- **열.** 열 관련 통계가 포함됩니다.

- **없음.** 모델을 스코어링하는 데 필요한 통계량만 포함됩니다.

## **z/OS용 IBM Db2 모델 - 회귀분석 트리 작성 옵션 - 트리 가지치기**

가지치기 옵션을 사용하여 회귀분석 트리의 가지치기 기준을 지정할 수 있습니다. 가지치기의 목적은 새 데이터에 대한 예상 정확도를 향상시키지 않는 지나치게 확장된 하위 그룹을 제거하여 과적합 위험을 줄이기 위한 것입니다.

**가지치기 측도.** 가지치기 측도를 사용하면 트리에서 리프를 제거한 후에 모델의 추정 정확도가 허용 가능한 한계내로 유지될 수 있습니다. 다음 측도 중 하나를 선택할 수 있습니다.

- **mse.** 평균 제곱 오차 - (기본값) 맞춤선이 데이터 점에 어느 정도 가까운지 측정합니다.
- **r2.** R-제곱 - 회귀 모형에 의해 설명되는 종속변수에서 편차의 비율을 측정합니다.
- **Pearson.** Pearson 상관 계수 - 정규적으로 분포된 선형 종속 변수 간의 관계 강도를 측정합니다.
- **Spearman.** Spearman 상관 계수 - Pearson 상관에 따라 약해 보이지만 실제로는 강할 수 있는 비선형 관계를 발견합니다.

**가지치기를 위한 데이터.** 학습 데이터의 일부 또는 모두를 사용하여 새 데이터에 대한 예상 정확도를 추정할 수 있습니다. 또는 이 용도로 지정된 테이블에서 별도의 가지치기 데이터 세트를 사용할 수 있습니다.

- **모든 학습 데이터 사용.** 이 옵션(기본값)은 모든 학습 데이터를 사용하여 모형 정확도를 추정합니다.
  - **가지치기를 위해 학습 데이터의 % 사용.** 가지치기 데이터에 대해 여기서 지정된 백분율을 사용하여 데이터를 두 개의 세트(학습에 대한 세트 하나와 가지치기에 대한 세트 하나)로 분할하려면 이 옵션을 사용하십시오.
- 난수 시드를 지정하여 스트림을 실행할 때마다 데이터가 동일한 방식으로 파티셔닝되도록 하려면 결과 복제를 선택하십시오. 가지치기에 사용되는 시드 필드에 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성 할 수 있습니다.
- **기존 테이블의 데이터 사용.** 모형 정확도 추정을 위한 별도의 가지치기 데이터 세트의 테이블 이름을 지정하십시오. 이 방법은 학습 데이터를 사용하는 것보다 신뢰성이 높은 것으로 간주됩니다.

## **z/OS용 IBM Db2 모델 - TwoStep**

TwoStep 노드는 큰 데이터 세트의 데이터를 군집화하는 방법을 제공하는 TwoStep 알고리즘을 구현합니다.

이 노드를 사용하면 사용 가능한 자원(예: 메모리 및 시간 제한조건)을 고려하면서 데이터를 군집화할 수 있습니다.

TwoStep 알고리즘은 다음과 같은 방법으로 데이터를 군집화하는 데이터베이스 마이닝 알고리즘입니다.

1. 군집화 기능(CF) 트리가 작성됩니다. 균형도가 높은 이 트리는 유사한 입력 레코드가 동일한 트리 노드의 일부가 되는 계층별 군집화를 위한 군집화 기능을 저장합니다.
2. CF 트리의 리프는 인메모리로 계층적으로 군집화되어 최종 군집 결과를 생성합니다. 최상의 군집 수는 자동으로 결정됩니다. 최대 군집 수를 지정하면 지정된 한계 내에서 최상의 군집 수가 결정됩니다.
3. 군집 결과는 K-평균 알고리즘과 유사한 알고리즘이 데이터에 적용되는 두 번째 단계에서 세분화됩니다.

## **z/OS용 IBM Db2 모델 - TwoStep 필드 옵션**

필드 옵션을 설정하여 업스트림 노드에 정의된 필드 역할 설정을 사용하도록 지정할 수 있습니다. 수동으로 필드 할당을 수행할 수도 있습니다.

**항목 선택.** 업스트림 소스 노드의 유형 템 또는 업스트림 유형 노드의 역할 설정을 사용하려면 이 옵션을 선택하십시오. 역할 설정은, 예를 들면, 목표 및 예측변수입니다.

**사용자 정의 필드 할당 사용.** 수동으로 목표, 예측변수 및 기타 역할을 할당하려면 이 옵션을 선택하십시오.

**필드.** 화살표를 사용하여 수동으로 이 목록의 항목을 화면의 오른쪽에 있는 역할 필드에 지정하십시오. 아이콘은 각 역할 필드에 대한 유효한 측정 수준을 나타냅니다.

**레코드 ID.** 고유 레코드 식별자로 사용할 필드입니다.

**예측변수(입력).** 예측에 대한 입력으로 하나 이상의 필드를 선택합니다.

## **z/OS용 IBM Db2 모델 - TwoStep 작성 옵션**

작성 옵션을 설정하여 고유의 목적에 맞게 모델 작성을 사용자 정의할 수 있습니다.

기본 옵션을 사용하여 모델을 작성하려면 **실행**을 클릭하십시오.

**거리 측도.** 이 매개변수는 데이터 점 사이의 거리를 측정하는 방법을 정의합니다. 거리가 멀수록 유사성이 떨어집니다. 옵션은 다음과 같습니다.

- **로그-우도.** 우도 측도는 변수에 확률 분포를 둡니다. 연속형 변수는 정규 분포로, 범주형 변수는 다항분포로 가정됩니다. 모든 변수를 독립변수로 가정합니다.

군집 수. 이 매개변수는 작성할 군집 수를 정의합니다. 옵션은 다음과 같습니다.

- 자동으로 군집 수 계산. 군집 수가 자동으로 계산됩니다. 최대값 필드에 군집 최대 수를 지정할 수 있습니다.
- 군집 수 지정. 작성할 군집 수를 지정하십시오.

통계량. 이 매개변수는 모델에 포함되는 통계 수를 정의합니다. 옵션은 다음과 같습니다.

- 모두. 모든 열 관련 통계량 및 모든 값 관련 통계량이 포함됩니다.

참고: 이 매개변수는 통계 최대 수를 포함하므로 시스템 성능에 영향을 줄 수도 있습니다. 모델을 그래픽 형식으로 보지 않으려면 없음을 지정하십시오.

- 열. 열 관련 통계가 포함됩니다.
- 없음. 모델을 스코어링하는 데 필요한 통계량만 포함됩니다.

결과 복제. 분석을 복제하기 위한 난수 시드를 설정하려면 이 선택란을 선택하십시오. 정수를 지정하거나 생성을 클릭하여 의사 난수 정수를 작성할 수 있습니다.

## **z/OS용 IBM Db2 모델 - TwoStep 너깃 - 모델 탭**

모델 탭에는 군집 필드에 대한 분포 및 요약 통계를 표시하는 다양한 그래픽 보기가 있습니다. 모델에서 데이터를 내보내거나 보기를 그래픽으로 내보낼 수 있습니다.

## **IBM Db2 for z/OS 모델 관리**

z/OS용 Db2 모델은 다른 IBM SPSS Modeler 모델과 동일한 방식으로 캔버스 및 모델 팔레트에 추가되며 거의 동일한 방식으로 사용할 수 있습니다.

z/OS용 Db2에서 직접 데이터를 스코어링하려면 다음 단계를 수행하십시오.

- 데이터가 위치한 z/OS용 Db2 데이터베이스에 SPSS Scoring Adapter를 설치하십시오.
- 데이터가 위치한 z/OS용 Db2 데이터베이스에 스트림이 연결되게 하십시오.

## **IBM Db2 for z/OS 모델 스코어링**

모델은 금색 모델 너깃 아이콘에 의해 캔버스에 표시됩니다. 너깃의 주 용도는 데이터를 스코어링하여 예측을 생성하거나 모델 특성의 추가 분석을 허용하는 것입니다. 나중에 이 절에서 설명되는 대로 테이블 노드를 너깃에 연결하고 스트림의 해당 분기를 실행하여 표시될 수 있는 하나 이상의 추가 데이터 필드 형식으로 스코어가 추가됩니다. 의사결정 트리 또는 회귀분석 트리에 대한 대화 상자 등의 일부 너깃 대화 상자에는 모델의 시각적 표시를 제공하는 모델 탭이 추가로 제공됩니다.

추가 필드는 대상 필드의 이름에 추가된 \$<id>- 접두부에 의해 구별됩니다. 여기서 <id>는 모델에 따라 다르며 추가 중인 정보의 유형을 식별합니다. 각각의 모델 너깃에 대한 주제에 다양한 식별자가 설명되어 있습니다.

스코어를 보려면 다음의 단계를 완료하십시오.

- 테이블 노드를 모델 너깃에 연결하십시오.
- 테이블 노드를 여십시오.
- 실행을 클릭하십시오.
- 테이블 출력 창의 오른쪽으로 스크롤하여 추가 필드 및 해당 스코어를 보십시오.

참고: 스코어링 프로세스는 액셀러레이터에서 실행되지 않고 Db2에서 실행되므로 스코어링의 입력 테이블은 물리적으로 Db2에 위치해야 합니다. 따라서 스코어링 입력으로는 Db2 기반 테이블 또는 가속화된 테이블만 사용될 수 있습니다. 스트림이 액셀러레이터 전용 테이블을 사용하는 경우 "THE STATEMENT CANNOT BE EXECUTED BY DB2 OR IN THE ACCELERATOR"와 같은 오류가 발생합니다.

## **z/OS용 IBM Db2 의사결정 트리 모델 너깃**

의사결정 트리 모델 너깃은 모델링 작업의 출력을 표시하며 사용자가 모델 스코어링에 대한 일부 옵션을 설정할 수 있도록 해줍니다.

의사결정 트리 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 두 개의 새 필드를 추가합니다.

표 26. 의사결정 트리의 모델 스코어링 필드	
추가되는 필드의 이름	의미
$\$I\text{-}target\_name$	현재 레코드의 예측값입니다.
$\$IP\text{-}target\_name$	예측 신뢰도(0.0 - 1.0)입니다.

참고: z/OS용 Db2의 제한사항으로 인해 열 이름은 잘릴 수 있습니다.

### **z/OS용 IBM Db2 의사결정 트리 너깃 - 모델 탭**

모델 탭은 의사결정 트리 모형의 예측변수 중요도를 그래픽 형식으로 표시합니다. 막대의 길이는 예측변수의 중요도를 나타냅니다.

### **z/OS용 IBM Db2 의사결정 트리 너깃 - 뷰어 탭**

뷰어 탭은 SPSS Modeler가 해당 의사결정 트리 모형의 트리 프리젠테이션을 표시하는 것과 동일한 방식으로 트리 모형의 트리 프리젠테이션을 표시합니다.

### **z/OS용 IBM Db2 K-평균 모델 너깃**

K-평균 모델 너깃은 학습 데이터 및 추정 프로세스에 대한 정보와 함께, 군집 모델에서 캡처한 모든 정보를 포함합니다.

K-평균 모델 너깃을 포함하는 스트림을 실행하는 경우, 노드가 소속군집 및 해당 레코드에 대해 지정된 군집 중심으로부터의 거리를 포함하는 두 개의 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생되는데, 소속군집의 경우  $\$KM\text{-}i$  앞에 추가되고 군집 중심의 경우  $\$KMD\text{-}g$  앞에 추가됩니다. 예를 들어, 모델 이름이 Kmeans인 경우, 새 필드 이름이  $\$KM\text{-}Kmeans$  및  $\$KMD\text{-}Kmeans$ 가 됩니다.

참고: z/OS용 Db2의 제한사항으로 인해 열 이름은 잘릴 수 있습니다.

### **z/OS용 IBM Db2 K-평균 너깃 - 모델 탭**

모델 탭에는 군집 필드에 대한 분포 및 요약 통계를 표시하는 다양한 그래픽 보기があります. 모델에서 데이터를 내보내거나 보기를 그래픽으로 내보낼 수 있습니다.

### **z/OS용 IBM Db2 Naive Bayes 모델 너깃**

Naive Bayes 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 두 개의 새 필드를 추가합니다.

표 27. Naive Bayes에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
$\$I\text{-}target\_name$	현재 레코드의 예측값입니다.
$\$IP\text{-}target\_name$	예측 신뢰도(0.0 - 1.0)입니다.

참고: z/OS용 Db2의 제한사항으로 인해 열 이름은 잘릴 수 있습니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

### **z/OS용 IBM Db2 회귀분석 트리 모델 너깃**

회귀분석 트리 모델 너깃을 포함한 스트림을 실행하면 노드는 목표 이름에서 이름이 파생된 두 개의 새 필드를 추가합니다.

표 28. 회귀분석 트리에 대한 모델 스코어링 필드	
추가되는 필드의 이름	의미
\$I-target_name	현재 레코드의 예측값입니다.
\$IS-target_name	예측값의 추정 표준 편차입니다.

참고: z/OS용 Db2의 제한사항으로 인해 열 이름은 잘릴 수 있습니다.

모델 너깃에 테이블 노드를 연결하고 테이블 노드를 실행하여 추가 필드를 볼 수 있습니다.

### **z/OS용 IBM Db2 회귀분석 트리 너깃 - 모델 탭**

모델 탭은 회귀분석 트리 모델의 예측변수 중요도를 그래픽 형식으로 표시합니다. 막대의 길이는 예측변수의 중요도를 나타냅니다.

### **z/OS용 IBM Db2 회귀분석 트리 너깃 - 뷰어 탭**

뷰어 탭은 SPSS Modeler가 해당 회귀분석 트리 모델의 트리 프리젠테이션을 표시하는 것과 동일한 방식으로 트리 모델의 트리 프리젠테이션을 표시합니다.

### **z/OS용 IBM Db2 TwoStep 모델 너깃**

TwoStep 모델 너깃을 포함하는 스트림을 실행하는 경우 노드는 해당 레코드의 지정된 군집 중심과의 거리 및 소속군집을 포함하는 2개의 새 필드를 추가합니다. 새 필드 이름은 모델 이름에서 파생되는데, 소속군집의 경우 \$TS-이 앞에 추가되고 군집 중심의 경우 \$TSD-가 앞에 추가됩니다. 예를 들어, 모델 이름이 MDL인 경우, 새 필드 이름은 \$TS-MDL 및 \$TSD-MDL이 됩니다.

## 주의사항

이 정보는 미국에서 제공되는 제품 및 서비스용으로 작성된 것입니다. 이 자료는 IBM에서 다른 언어로 사용 가능합니다. 그러나 자료에 접근하기 위해서는 해당 언어로 된 제품 또는 제품 버전의 사본이 필요할 수 있습니다.

IBM은 다른 국가에서 이 책에 기술된 제품, 서비스 또는 기능을 제공하지 않을 수도 있습니다. 현재 사용할 수 있는 제품 및 서비스에 대한 정보는 한국 IBM 담당자에게 문의하십시오. 이 책에서 IBM 제품, 프로그램 또는 서비스를 언급했다고 해서 해당 IBM 제품, 프로그램 또는 서비스만을 사용할 수 있다는 것을 의미하지는 않습니다. IBM의 지적 재산권을 침해하지 않는 한, 기능상으로 동등한 제품, 프로그램 또는 서비스를 대신 사용할 수도 있습니다. 그러나 비IBM 제품, 프로그램 또는 서비스의 운영에 대한 평가 및 검증은 사용자의 책임입니다.

IBM은 이 책에서 다루고 있는 특정 내용에 대해 특허를 보유하고 있거나 현재 특허 출원 중일 수 있습니다. 이 책을 제공한다고 해서 특허에 대한 라이센스까지 부여하는 것은 아닙니다. 라이센스에 대한 의문사항은 다음으로 문의하십시오.

### 07326

서울특별시 영등포구  
국제금융로 10, 3IFC  
한국 아이.비.엠 주식회사  
대표전화서비스: 02-3781-7114

2바이트(DBCS) 정보에 관한 라이센스 문의는 한국 IBM에 문의하거나 다음 주소로 서면 문의하시기 바랍니다.

*Intellectual Property Licensing  
Legal and Intellectual Property Law  
IBM Japan Ltd.  
19-21, Nihonbashi-Hakozakicho, Chuo-ku  
Tokyo 103-8510, Japan*

IBM은 타인의 권리 비침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 이 책을 "현상태대로" 제공합니다. 일부 국가에서는 특정 거래에서 명시적 또는 묵시적 보증의 면책사항을 허용하지 않으므로, 이 사항이 적용되지 않을 수도 있습니다.

이 정보에는 기술적으로 부정확한 내용이나 인쇄상의 오류가 있을 수 있습니다. 이 정보는 주기적으로 변경되며, 변경된 사항은 최신판에 통합됩니다. IBM은 이 책에서 설명한 제품 및/또는 프로그램을 사전 통지 없이 언제든지 개선 및/또는 변경할 수 있습니다.

이 정보에서 언급되는 비IBM 웹 사이트는 단지 편의상 제공된 것으로, 어떤 방식으로든 이들 웹 사이트를 옹호하고자 하는 것은 아닙니다. 해당 웹 사이트의 자료는 본 IBM 제품 자료의 일부가 아니므로 해당 웹 사이트 사용으로 인한 위험은 사용자 본인이 감수해야 합니다.

IBM은 귀하의 권리를 침해하지 않는 범위 내에서 적절하다고 생각하는 방식으로 귀하가 제공한 정보를 사용하거나 배포할 수 있습니다.

(i) 독립적으로 작성된 프로그램과 기타 프로그램(본 프로그램 포함) 간의 정보 교환 및 (ii) 교환된 정보의 상호 이용을 목적으로 본 프로그램에 관한 정보를 얻고자 하는 라이센스 사용자는 다음 주소로 문의하십시오.

### 07326

서울특별시 영등포구  
국제금융로 10, 3IFC  
한국 아이.비.엠 주식회사  
대표전화서비스: 02-3781-7114

이러한 정보는 해당 조건(예를 들면, 사용료 지불 등)하에서 사용될 수 있습니다.

이 정보에 기술된 라이센스가 부여된 프로그램 및 프로그램에 대해 사용 가능한 모든 라이센스가 부여된 자료는 IBM이 IBM 기본 계약, IBM 프로그램 라이센스 계약(IPLA) 또는 이와 동등한 계약에 따라 제공한 것입니다.

인용된 성능 데이터와 고객 예제는 예시 용도로만 제공됩니다. 실제 성능 결과는 특정 구성과 운영 조건에 따라 다를 수 있습니다.

비IBM 제품에 관한 정보는 해당 제품의 공급업체, 공개 자료 또는 기타 범용 소스로부터 얻은 것입니다. IBM에서는 이러한 제품들을 테스트하지 않았으므로, 비IBM 제품과 관련된 성능의 정확성, 호환성 또는 기타 청구에 대해서는 확신할 수 없습니다. 비IBM 제품의 성능에 대한 의문사항은 해당 제품의 공급업체에 문의하십시오.

IBM이 제시하는 방향 또는 의도에 관한 모든 언급은 특별한 통지 없이 변경될 수 있습니다.

이 정보에는 일상의 비즈니스 운영에서 사용되는 자료 및 보고서에 대한 예제가 들어 있습니다. 이들 예제에는 개념을 가능한 완벽하게 설명하기 위하여 개인, 회사, 상표 및 제품의 이름이 사용될 수 있습니다. 이들 이름은 모두 가공의 것이며 실제 인물 또는 기업의 이름과 유사하더라도 이는 전적으로 우연입니다.

## 상표

---

IBM, IBM 로고 및 ibm.com은 전세계 여러 국가에 등록된 International Business Machines Corp.의 상표 또는 등록상표입니다. 기타 제품 및 서비스 이름은 IBM 또는 타사의 상표입니다. 현재 IBM 상표 목록은 웹 "저작권 및 상표 정보"([www.ibm.com/legal/copytrade.shtml](http://www.ibm.com/legal/copytrade.shtml))에 있습니다.

Adobe, Adobe 로고, PostScript 및 PostScript 로고는 미국 및/또는 기타 국가에서 사용되는 Adobe Systems Incorporated의 등록상표 또는 상표입니다.

Intel, Intel 로고, Intel Inside, Intel Inside 로고, Intel Centrino, Intel Centrino 로고, Celeron, Intel Xeon, Intel SpeedStep, Itanium 및 Pentium은 미국 또는 기타 국가에서 사용되는 Intel Corporation 또는 그 계열사의 상표 또는 등록상표입니다.

Linux는 미국 또는 기타 국가에서 사용되는 Linus Torvalds의 등록상표입니다.

Microsoft, Windows, Windows NT 및 Windows 로고는 미국 또는 기타 국가에서 사용되는 Microsoft Corporation의 상표입니다.

UNIX는 미국 및 기타 국가에서 사용되는 The Open Group의 등록상표입니다.

Java 및 모든 Java 기반 상표와 로고는 Oracle 및/또는 그 계열사의 상표 또는 등록상표입니다.

## 제품 문서의 이용 약관

---

다음 이용 약관에 따라 이 책을 사용할 수 있습니다.

### 적용성

본 이용 약관은 IBM 웹 사이트의 모든 이용 약관에 추가됩니다.

### 개인적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 개인적, 비상업적 용도로 복제할 수 있습니다. 귀하는 IBM의 명시적 동의 없이 본 발행물 또는 그 일부를 배포 또는 전시하거나 2차적 저작물을 만들 수 없습니다.

### 상업적 사용

모든 소유권 사항을 표시하는 경우에 한하여 귀하는 이 책을 귀하 기업집단 내에서만 복제, 배포 및 전시할 수 있습니다. 귀하는 귀하의 기업집단 외에서는 IBM의 명시적 동의 없이 이 책의 2차적 저작물을 만들거나 이 책 또는 그 일부를 복제, 배포 또는 전시할 수 없습니다.

### 권한

본 허가에서 명시적으로 부여된 경우를 제외하고, 이 책이나 이 책에 포함된 정보, 데이터, 소프트웨어 또는 기타 지적 재산권에 대한 어떠한 허가나 라이센스 또는 권한도 명시적 또는 묵시적으로 부여되지 않습니다.

IBM은 이 책의 사용이 IBM의 이익을 해친다고 판단되거나 위에서 언급된 지시사항이 준수되지 않는다고 판단하는 경우 언제든지 부여한 허가를 철회할 수 있습니다.

귀하는 미국 수출법 및 관련 규정을 포함하여 모든 적용 가능한 법률 및 규정을 철저히 준수하는 경우에만 본 정보를 다운로드, 송신 또는 재송신할 수 있습니다.

IBM은 이 책의 내용과 관련하여 아무런 보장을 하지 않습니다. 타인의 권리 침해, 상품성 및 특정 목적에의 적합성에 대한 묵시적 보증을 포함하여(단, 이에 한하지 않음) 묵시적이든 명시적이든 어떠한 종류의 보증 없이 현 상태대로 제공합니다.



# 색인

## A

Analysis Services

모델 관리 12

예제 20

의사결정 트리 20

Apriori

Microsoft 15

Oracle 데이터 마이닝 34, 35

ARIMA 모델

IBM Netezza Analytics 57, 61

## D

distance 함수

Oracle K-평균 33

DSN

구성 11

## G

Gaussian kernel

Oracle 지원 벡터 머신 28

## I

IBM

모델 관리 46

IBM Netezza Analytics

모델 관리 64

모델 옵션 46

분열 군집 48

분열 군집 모델 너깃 68

분열 군집 작성 옵션 49

분열 군집 필드 옵션 48

선형 회귀 54

선형 회귀 모형 너깃 70

선형 회귀 작성 옵션 54

시계열 57

시계열 모델 너깃 70, 71

시계열 모델 옵션 62

시계열 작성 옵션 60, 61

시계열 필드 옵션 60

의사결정 트리 51

의사결정 트리 모형 너깃 64, 65, 70

의사결정 트리 작성 옵션 53

의사결정 트리 필드 옵션 52

일반화 선형 49

일반화 선형 모형 너깃 49, 71, 72

일반화 선형 모형 옵션 50

최근접 이웃(KNN) 54

필드 옵션 45

회귀분석 트리 47

회귀분석 트리 모델 너깃 69, 70

회귀분석 트리 작성 옵션 47

Bayes Net 57

IBM Netezza Analytics (계속)

Bayes Net 모델 너깃 66

Bayes Net 작성 옵션 57

Bayes Net 필드 옵션 57

IBM SPSS Modeler에서 구성 41, 42, 44, 46

K-평균 56

K-평균 모델 너깃 65, 66

K-평균 작성 옵션 56

K-평균 필드 옵션 56

KNN 모델 너깃 67, 68

KNN 모형 옵션 55

Naive Bayes 57

Naive Bayes 모델 너깃 67

PCA 63

PCA 모델 너깃 69

PCA 작성 옵션 63

PCA 필드 옵션 63

TwoStep 62

TwoStep 모델 너깃 72

TwoStep 작성 옵션 62

TwoStep 필드 옵션 62

IBM SPSS Modeler

데이터베이스 마이닝 5

문서 3

IBM SPSS Modeler Server 1

IBM SPSS Modeler Solution Publisher

Oracle Data Mining 모델 25

In-Database 모델링 18

## K

K-평균

IBM Netezza Analytics 56, 65, 66

Oracle 데이터 마이닝 32, 33

z/OS용 IBM Db2 77, 83

KNN 모델

IBM Netezza Analytics 67, 68

## M

MDL 27

Microsoft

군집화 모델링 9, 11, 18

로지스틱 회귀분석 9

로지스틱 회귀분석 모델링 11, 18

모델 관리 12

선형 회귀 9

선형 회귀 모델링 11, 18

시퀀스 군집화 9

신경망 9

신경망 모델링 11, 18

연관 규칙 모델링 9, 11, 18

의사결정 트리 모델링 9, 11, 18

Analysis Services 9, 11, 18

Naive Bayes 모델링 9, 11, 18

Microsoft Analysis Services 18-20

## N

Naive Bayes  
고급 옵션 15  
모델 옵션 14  
서버 옵션 14  
스코어링 - 서버 옵션 18  
스코어링 - 요약 옵션 18  
IBM Netezza Analytics 57, 67  
Oracle 데이터 마이닝 26, 27  
z/OS용 IBM Db2 78, 83

Naive Bayes 모델  
적응형 Oracle Bayes 네트워크 27  
IBM Netezza Analytics 67

Netezza  
모델 관리 46

Netezza 트리 모델의 리프 51, 78  
Netezza 트리 모델의 인스턴스 가중치 51  
Netezza 트리 모델의 클래스 가중치 51  
Netezza 트리 모델의 클래스 레이블 51, 78  
NMF  
Oracle 데이터 마이닝 33, 34

## O

O-군집  
Oracle 데이터 마이닝 32

ODBC  
구성 11  
IBM Netezza Analytics에 대한 구성 41, 42, 44, 46  
Oracle에 대한 구성 23-26  
SQL Server 구성 11  
z/OS용 IBM Db2 구성 77

ODM. Oracle 데이터 마이닝 참조 23

Oracle 데이터 마이닝  
데이터 준비 38  
모델 관리 37, 38  
속성 중요도(AI) 36  
예제 39, 40  
오분류 비용 37  
의사결정 트리 31  
일반화 선형 모형(GLM) 30, 31  
일치도 검사 37  
적응형 Bayes 네트워크 27, 28  
지원 벡터 머신 28, 29  
최소 설명 길이(MDL) 35, 36  
Apriori 34, 35  
IBM SPSS Modeler에서 구성 23-26  
K-평균 32, 33  
Naive Bayes 26, 27  
NMF 33, 34  
O-군집 32

Oracle Data Miner 38

## P

PCA 모델  
IBM Netezza Analytics 63, 69

## S

SID  
Oracle 연결 24

## Solution Publisher

Oracle Data Mining 모델 25  
SQL 생성 6  
SQL Server  
구성 11  
ODBC 연결 11  
SVM. 지원 벡터 머신 참조 28

## T

tnsnames.ora 파일 24  
TwoStep  
IBM Netezza Analytics 62, 72  
z/OS용 IBM Db2 81, 82, 84

## Z

Z 점수  
데이터 표준화 28, 38  
z/OS용 Db2 모델링  
z/OS용 IBM Db2 73, 75, 77  
z/OS용 IBM Db2  
모델 옵션 77  
의사결정 트리 78  
의사결정 트리 모형 너깃 82-84  
의사결정 트리 작성 옵션 79  
의사결정 트리 필드 옵션 78  
필드 옵션 76  
회귀분석 트리 80  
회귀분석 트리 모델 너깃 83, 84  
회귀분석 트리 작성 옵션 80  
IBM SPSS Modeler에서 구성 75, 77  
K-평균 77  
K-평균 모델 너깃 83  
K-평균 작성 옵션 77  
K-평균 필드 옵션 77  
Naive Bayes 78  
Naive Bayes 모델 너깃 83  
TwoStep 81  
TwoStep 모델 너깃 82, 84  
TwoStep 작성 옵션 81  
TwoStep 필드 옵션 81  
z/OS용 IBM Db2 모델 관리 82  
z/OS용 IBM Db2 및 z/OS용 IBM Analytics Accelerator 구성 74  
z/OS용 IBM Db2와의 통합 73  
z/OS용 IBM Db2와의 통합을 위한 요구사항 73

## 가

가지치기된 Naive Bayes 모델  
적응형 Oracle Bayes 네트워크 27  
값의 보간법, IBM Netezza Analytics 시계열 58  
계절 추세 분해, IBM Netezza Analytics 57  
고유 필드  
적응형 Oracle Bayes 네트워크 27  
Oracle 데이터 마이닝 25  
Oracle 지원 벡터 머신 28  
Oracle Apriori 31, 35  
Oracle K-평균 33  
Oracle MDL 36  
Oracle Naive Bayes 26  
Oracle NMF 33

- 고유 필드 (계속)
  - Oracle O-Cluster 32
- 교차 검증
  - Oracle Naive Bayes 26
- 구성
  - z/OS용 IBM Db2 및 z/OS용 IBM Analytics Accelerator 74
- 군집 수
  - Oracle K-평균 33
  - Oracle O-Cluster 32
- 군집화
  - 고급 옵션 14
  - 모델 옵션 14
  - 서버 옵션 14
  - 스코어링 - 서버 옵션 18
  - 스코어링 - 요약 옵션 18
  - IBM Netezza Analytics 68
- 나**
- 내보내기
  - Analysis Services 모델 20
- 노드
  - 생성 20
  - 노드 생성 20
- 다**
- 다기능 모델
  - 적응형 Oracle Bayes 네트워크 27
- 단일 필드 모델
  - 적응형 Oracle Bayes 네트워크 27
- 대응별 임계값
  - Oracle Naive Bayes 27
- 데이터 검토 노드 21, 40
- 데이터 구간화
  - Oracle 모델 38
- 데이터 표준화
  - Oracle 모델 38
- 데이터베이스
  - In-Database 모델링 6, 9, 11, 12, 18
- 데이터베이스 마이닝
  - 구성 11
  - 데이터 준비 6
  - 모델 작성 6
  - 예 20
  - 최적화 옵션 6
  - IBM SPSS Modeler 사용 5
- 데이터베이스 모델링
  - IBM Netezza Analytics 41, 42, 44, 46
  - Oracle 23–26
- 라**
- 로지스틱 회귀분석
  - 고급 옵션 15
  - 모델 옵션 14
  - 서버 옵션 14
  - 스코어링 - 서버 옵션 18
  - 스코어링 - 요약 옵션 18
- 마**
- 모델
- 모델 (계속)
  - 내보내기 6
  - 일관성 문제 7
  - 저장 6
  - 평가 21, 40
  - Analysis Services 관리 12
  - In-Database 모델 스코어링 6
  - In-Database 모델 작성 6
  - Netezza 관리 46
  - Netezza 나열 46
  - Oracle 찾아보기 27
- 모델 너깃
  - IBM Netezza Analytics 49, 64–72
  - z/OS용 IBM Db2 82–84
- 모델 옵션
  - IBM Netezza Analytics 46, 50, 55, 62
  - z/OS용 IBM Db2 77
- 모델링 노드
  - In-Database 모델링 6, 9, 11, 12, 18
  - Microsoft 군집화 12
  - Microsoft 로지스틱 회귀분석 12
  - Microsoft 선형 회귀 12
  - Microsoft 시계열 12
  - Microsoft 시퀀스 군집화 12
  - Microsoft 신경망 12
  - Microsoft 연관 규칙 12
  - Microsoft 의사결정 트리 12
  - Microsoft Naive Bayes 12
- 문서 3
- 바**
- 발행자 노드
  - Oracle Data Mining 모델 25
- 배포 21, 40
- 베이지안 신경망 모형
  - IBM Netezza Analytics 57, 66
- 복잡도 요인
  - Oracle 지원 벡터 머신 29
- 복잡도 페널티 14–16
- 분열 군집
  - IBM Netezza Analytics 48, 49, 68
- 분할 기준
  - Oracle K-평균 33
- 불순도 메트릭
  - Oracle Apriori 31
- 불순도 측도
  - 의사결정 트리 79
  - Netezza 의사결정 트리 53
- 비용
  - Oracle 26
- 사**
- 사전 확률
  - Oracle 데이터 마이닝 29
- 서버
  - Analysis Services 실행 14, 18
- 선형 커널
  - Oracle 지원 벡터 머신 28
- 선형 회귀
  - 고급 옵션 15
  - 모델 옵션 14

선형 회귀 (계속)  
    서버 옵션 14  
    스코어링 - 서버 옵션 18  
    스코어링 - 요약 옵션 18  
    IBM Netezza Analytics 47, 54, 70  
    z/OS용 IBM Db2 80

속성 중요도(AI)  
    Oracle 데이터 마이닝 36

수렴허용  
    Oracle 지원 벡터 머신 29

스코어링 6, 64, 82  
스페셜리즘 분석, IBM Netezza Analytics 57  
시계열

    IBM Netezza Analytics 60–62  
시계열(IBM Netezza Analytics) 57, 70, 71

시계열(Microsoft)

    고급 옵션 16  
    모델 옵션 16  
    설정 옵션 16

시퀀스군집화

    모델 옵션 14

시퀀스군집화(Microsoft)  
    고급 옵션 17  
    필드 옵션 17

신경망

    고급 옵션 15  
    모델 옵션 14  
    서버 옵션 14  
    스코어링 - 서버 옵션 18  
    스코어링 - 요약 옵션 18

싱글톤 임계값

    Oracle Naive Bayes 27

## 아

애플리케이션 예제 3  
엔트로피 불순도 측도 53

엡실론

    Oracle 지원 벡터 머신 29

연관 규칙

    고급 옵션 15  
    모델 옵션 14  
    서버 옵션 14  
    스코어링 - 서버 옵션 18  
    스코어링 - 요약 옵션 18

연관 규칙 모델

    Microsoft 15

예제

    개요 3  
    데이터베이스 마이닝 20, 21, 40  
    애플리케이션 안내서 3

오분류 비용

    Oracle 26

요구사항

    z/OS용 IBM Db2 73

의사결정 트리

    고급 옵션 14  
    모델 옵션 14  
    서버 옵션 14  
    스코어링 - 서버 옵션 18  
    스코어링 - 요약 옵션 18

IBM Netezza Analytics 51–53, 64, 65, 70

Microsoft Analysis Services 9, 11, 18

    Oracle 데이터 마이닝 31

의사결정 트리 (계속)  
    z/OS용 IBM Db2 78, 79, 82–84  
일반화 선형 모델  
    IBM Netezza Analytics 49–51, 71, 72  
일반화 선형 모형(GLM)  
    Oracle 데이터 마이닝 30, 31

## 자

작성 옵션  
    IBM Netezza Analytics 47, 49, 53, 54, 56, 57, 60–62  
    z/OS용 IBM Db2 77, 79–81  
적응형 Bayes 네트워크  
    Oracle 데이터 마이닝 27, 28  
정규화 방법  
    Oracle 지원 벡터 머신 28  
    Oracle K-평균 33  
    Oracle NMF 33  
    지니 불순도 측도 53  
    지수평활  
        IBM Netezza Analytics 57  
    지원 벡터 머신  
        Oracle 데이터 마이닝 28, 29

## 차

최근접 이웃 모델  
    IBM Netezza Analytics 54, 55, 67, 68  
최소 설명 길이 27  
최소 설명 길이(MDL)  
    Oracle 데이터 마이닝 35, 36  
최소-최대  
    데이터 표준화 28, 38

## 카

키  
    모델 키 7

## 타

탐색 21, 40

## 파

파티션 데이터 34  
파티션 필드  
    선택 34  
평가 21, 40  
포트  
    Oracle 연결 24  
표준 편차  
    Oracle 지원 벡터 머신 29  
필드 옵션  
    IBM Netezza Analytics 45, 48, 52, 56, 57, 60, 62, 63  
    z/OS용 IBM Db2 76–78, 81

## 하

호스트 이름  
    Oracle 연결 24  
회귀분석 트리

회귀분석 트리 (계속)

IBM Netezza Analytics 47, 69, 70

z/OS용 IBM Db2 80, 83, 84





**IBM.**<sup>®</sup>