T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KARAR AĞAÇLARI KULLANILARAK TRAFİK KAZALARININ NEDENLERİNİN ARAŞTIRILMASI: SAKARYA İLİ VAKA ÇALIŞMASI

YÜKSEK LİSANS TEZİ

Joseph Doucke OMENDE KAHUDI

Enstitü Anabilim Dalı	**:**	**İNŞAAT MÜHENDİSLİĞİ**
Enstitü Bilim Dalı	**:**	**ULAŞTIRMA**
Tez Danışmanı	**:**	**Dr. Öğr.Üyesi. HAKAN ASLAN**

Ağustos 2021

T.C.
SAKARYA ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

KARAR AĞAÇLARI KULLANILARAK TRAFİK KAZALARININ NEDENLERİNİN ARAŞTIRILMASI: SAKARYA İLİ VAKA ÇALIŞMASI

YÜKSEK LİSANS TEZİ

Joseph Doucke OMENDE KAHUDI

Enstitü Anabilim Dalı	**:**	**İNŞAAT MÜHENDİSLİĞİ**
Enstitü Bilim Dalı	**:**	**ULAŞTIRMA**

Bu tez 28/08/2021 tarihinde aşağıdaki jüri tarafından oybirliği / oyçokluğu ile kabul edilmiştir.

Jüri Başkanı	**Üye**	**Üye**

BEYAN

Tez içindeki tüm verilerin akademik kurallar çerçevesinde tarafımdan elde edildiğini, görsel ve yazılı tüm bilgi ve sonuçların akademik ve etik kurallara uygun şekilde sunulduğunu, kullanılan verilerde her hangi bir tahrifat yapılmadığını, başkalarının eserlerinden yararlanılması durumunda bilimsel normlara uygun olarak atıfta bulunulduğunu, tezde yer alan verilerin bu üniversite veya başka bir tez çalışmasında kullanılmadığını beyan ederim.

Joseph Doucke OMENDE KAHUDI

28.08.2021

TEŞEKKÜR

Onsuz varlığımızın anlamsız olacağı tüm yaşamın ışığı ve kaynağı olan TANRI'ımıza karşı minnet duygularımı ifade etmeyi bir borç bilirim.

Bu tez çalışmamda akademik danışmanlığımı yapan Dr. Öğr.Üyesi Hakan ASLAN'a teşekkür etmek isterim. Tez yazım sürecinde yardımcı oldukları için Dr. Üğr. Üyesi Caner ERDEN ve Arş. Gör. Zeliha Çağla KUYUMCU'ya da teşekkür etmek istiyorum. Sakarya Üniversitesi Fen Bilimleri Enstitüsü'nün personellerine de ayrıca teşekkürlerimi sunuyorum.

Türkiye'de lisans üstü çalışma yapabilmeme finansal destek sağlayarak bu imkânı bana sunduğu için YTB'ye de çok teşekkür etmek isterim.

Eğitimime başladığımdan beri aileme destekleri için teşekkür ederim.

Sevgili dostlarım ve meslektaşlarım, desteğiniz için teşekkür ederim. Bazıları için teşekkür edecek kelimeler bulamıyorum. Sizlere bütün kalbimle söylüyorum, çok teşekkür ederim.

İÇİNDEKİLER

SİMGELER VE KISALTMALAR LİSTESİ

ARFF	: Attribute Relation File Forma
CART	: Classification And Regression Trees
CHAID	: Chi-squared Automatic Interaction Detector
CLI	: Command-Line Interface
CSV	: Comma – Separated Values
JDBC	: Java Database Connectivity
KA	: Karar Ağaçları
ML	: Machine Learning
MPP	: Massiely Parallel Processors
NCR	: National Cash Register
SQL	: Structured Query Language
URL	: Uniform Resource Locator
VM	: Veri Madenciliği
VS	: Veri Setini

ŞEKİLLER LİSTESİ

TABLOLAR LİSTESİ

ÖZET

Anahtar kelimeler: Trafik kazaları, Veri madenciliği, WEKA ve Karar ağaçları

Türkiye'de trafik kazalarından kaynaklanan ölüm ve yaralanma oranlarında son zamanlarda iyileştirmeler yapılmış olsa da trafik ve yol güvenliği toplumsal yaşamı derinden etkilemeye devam eden bir konu olmaya devam etmektedir. Bu çalışmada Türkiye'de Sakarya ilindeki trafik kazaları incelenmiştir. Bu çalışmanın verileri Sakarya İl Emniyet Müdürlüğü'nden elde edilmiştir. Veri setinde meydana gelen kazalarla ilgili 139 özellik sınıfı bulunmaktadır. Bu çalışmada öncelikle veri setini sınıflandırmak için veri setinin önemli özellikleri ortaya konmuştur. Daha sonra, sınıflandırma amaçlarına ulaşmak ve veri kümesinden daha kesin bilgi ve ilişkileri keşfetmek için literatürde sıklıkla kullanılan karar ağaçları sınıflandırma algoritması, WEKA yazılımı çerçevesinde kullanıldı. Bu anlamda J48 ve RepTrees algoritmalarının performans değerlendirmesi, sonuçların gösterdiği sınıflandırma kalitesine göre yapılmıştır.

INVESTIGATING THE CAUSES OF TRAFFIC ACCIDENTS USING DECISION TREES: CITY OF SAKARYA CASE STUDY

SUMMARY

Keywords: Traffic accidents, Data mining, WEKA and Decision trees

Although improvements have been made recently in the death and injury rates resulting from traffic accidents in Turkey, traffic and road safety continues to be an issue that continues to affect social life deeply. In this study, traffic accidents in Sakarya province in Turkey were investigated. The data of this study was obtained from Sakarya Provincial Police Department. There are 139 feature classes related to the accident that occurred in the dataset. In this study, first of all, important features of the data set are revealed to classify the data set. Second, the classification algorithm frequently used in the literature to achieve classification purposes; We used WEKA software to make decision trees that are used to discover more precise information and relationships from the dataset. The performance evaluation of the algorithms was evaluated according to the classification quality shown by the results.

BÖLÜM 1. GİRİŞ

Veri madenciliği; istatistik, yapay zeka veya bilgisayar bilimi gibi çeşitli bilimsel disiplinlerden bir dizi algoritma kullanmayı, verilerden modeller oluşturmayı, yani ilginç yapılar veya ön kriterlere göre örüntüler bulmayı ve mümkün olduğunca fazla bilgi üretmeyi sağlamaktadır. Günümüz uzmanlarına göre en değerli kaynak olarak petrol değil, veri görülmektedir. Bir diğer ifade ile verilere dijital çağın petrol`ü olarak bakılmaktadır. Büyük veri analizlerinin insanların yaşama biçimi, toplumlar ve bilimsel analiz süreçleri üzerinde etkileri gün geçtikçe artmaktadır. Bu anlamda bakıldığında dijital veriler; analistlerin, ekonomistlerin, mühendislerin ve bilim insanlarının çeşitli değişkenler arasındaki ilişkileri anlamalarını ve doğru modelleme tahminleri ile optimum amaç fonksiyonlarına ulaşabilmelerini sağlayabilmektedir. Veri analizleri, yapay zeka dahil olmak üzere farklı karmaşık bilimsel sistemleri kullanarak, alınacak stratejik kararların performansını artıracak büyük bir potansiyele sahip olabilmektedir.

Türkiye'nin en kalabalık bölgesi olan Marmara`da bulunan Sakarya'da trafik kazaları en önemli ölüm nedenlerinden biri olarak değerlendirilmektedir. Artan araç sayısına bağlı olarak trafik kazalarının da artması beklenmektedir. Sakarya'da trafiğe kayıtlı araç sayısı Mart 2020 sonu itibariyle 290.408 iken bu sayı 2021 yılı Mart ayı sonu itibarıyla 302.353'e ulaşmıştır. Rakamlardan da anlaşılacağı üzere Sakarya'da trafiğe katılan araç sayısı son bir yılda % 4,11 oranında artış göstermiştir. Sakarya ilinde trafik kazalarında meydana gelen yaralanmaların ciddiyetine ilişkin araştırmalar oldukça sınırlı olduğu için, karayolu güvenliği politikalarının etkinliği henüz net olarak belirlenememiştir. Bu araştırma, Sakarya'daki kaza sayılarını ve oluşan olumsuz etkilerini azaltmak için mevcut kaza verilerini analiz ederek, problematik yapının belirlenmesi ve gerekli stratejik önerilerin ortaya konulması amacıyla yapılmıştır.

1.1. Literatür

Trafik Kaza Analizlerinde (TKA) çok terimli ve sıralı logit regresyon gibi regresyon modelleri oluşturmak için uygulanan probit veya logit regresyon korelasyonları yaygın olarak kullanılan geleneksel yöntemlerdir [1].

Birçok vaka çalışmasında olumsuz hava koşullarının ölümler, yaralanmalar ve kaza sıklığı üzerindeki etkisinin analizinde, sıralı probit regresyonu kullanılmıştır [2].

Başka bir çalışmada, sürücü, araç ve yol özelliklerine ilişkin çeşitli parametrelerin yaralanma ve ölümlerin ciddiyeti üzerindeki etkisini tahmin etmek ve belirlemek için sıralı probit modeli uygulanmıştır [3].

Aynı veya potansiyel risk faktörlerinin etkisini anlamak için önceki çalışmalara benzer sıralı modeller kullanarak önemli araştırmalar yapılmıştır [4].

Alternatif analiz metodu olarak, çok terimli logit regresyon çalışmaları da birçok kaza sonucu yaralanma seviyesi belirleme çalışmalarında kullanılmıştır [5].

Karma logit modeli de karayolu kazalarını analiz etmek için uygulanan bir başka ayrık seçim modelidir [6].

Karışık logit modeli, normal bir dağılımla sınırlı olmadığı ve basit seçim olasılığı simülasyonlarına sahip olduğu için logit normunu aşabilir. Bu genellikle gereksiz etkilerin modellenmemesine ve varyantların homojenliği hakkındaki varsayımlardan kaçınılmasına izin verir. Birçok trafik kazası çalışması, yaralanma seviyesini tahmin etmek için karışık bir mantık uygulamıştır [7].

Diğer çalışmalar, bağımlı değişkenler arasındaki korelasyonu belirlemek için öngörücü bir yaklaşım olarak iç içe logiti seçmiştir. Tahmin ve sınıflandırma problemlerinde kullanılan bir diğer model ise karar ağacıdır. Karar ağacı temel olarak, özyinelemeli bisektör dahil olmak üzere çeşitli uygulamalardan oluşur; sınıflandırma

ve regresyon ağacı, C4.5; C5.0; ki-kare otomatik etkileşim dedektörü, M5P. Karar ağacı uygulamanın avantajlarından biri, bağımsız değişkenler ile bağımlı değişken arasında önceden tanımlanmış bir temel ilişkiye gerek duyulmamasıdır. Sonuç olarak, çeşitli araştırmacılar karayolu trafik kazası çalışmalarında karar ağacı modellerini uygulamışlardır. Bir araştırma yayaların, bisikletlilerin ve motosikletlerin en savunmasız yol kullanıcıları olduğunu doğruladı [8].

Bir başka çalışmada ise kaza tipi ve kaza nedeninin trafik kazalarının şiddetini etkileyen önemli parametreler olduğu tespit edilmiştir [9].

Kamyon kazaları için yapılan bir diğer araştırma, ciddi yaralanmaların artmasında diğer faktörlerin yanı sıra emniyet kemerlerinin, çarpışma tipinin ve kaza yerinin de önemli bir rol oynadığını göstermiştir [10].

Bayes modeli, trafik kazalarının ciddiyetini analiz etmede uygulanabilecek başka bir yaklaşımdır. Bayes modeli, ilgili önceki olasılıkların dahil edilmesini hesaba katan bir takım avantajlara sahip olduğundan, önceki olasılıkların bir sonucu olarak bir hipotezin doğru olma olasılığını belirler. Bayes modelini uygulamanın dezavantajı, ön tahminlerin zaman alıcı ve bazen hesaplama açısından çözümlenmemiş olmasıdır. Birkaç araştırmacı, Bayes Ağlarını kullanarak trafik travması üzerine çalışmalar yürütmüştür. Bu çalışmalardan biri, sürücünün yaşı, kaza türü ve aydınlatma koşullarının ciddi veya ölümcül yaralanmalarla ilişkili olduğunu göstermiştir [11].

Başka bir araştırmada, sürücünün yaşının ve emniyet kemeri kullanmamasının, diğer değişkenlerin yanı sıra yaralanma olasılığını artırabildiği belirtilmiştir [12].

Trafik kazalarını analiz etmek veya yaralanmanın ciddiyetini tahmin etmek için kullanılan bir diğer iyi bilinen makine öğrenimi modeli, destek vektör makineleridir. Önceki modellere benzer şekilde, vektör makine modellerinin avantajları ve dezavantajları vardır. Avantajları, yarı yapılandırılmış ve yapılandırılmamış verilerle yapılan analizlerde iyi performans göstermeleridir. Uygun bir çekirdek işlevinin kolayca seçilememesi ve nihai modelin yorumlanması için çaba gerektirmesi ise

olumsuz tarafları olarak belirtilebilir. Araştırmacılar trafik kazalarının ciddiyet analizlerine ilişkin çeşitli çalışmalarda, kaza ciddiyet tahminlerini elde etmek veya parametrelerle ilgili hassas karşılaştırmaları yapmak için bir destek vektör makinesi kullanmışlardır.

Sakarya ilinde trafik kazalarında meydana gelen ölümlerin yüksek olması, bu ölümleri azaltmaya yönelik bir çalışmanın gerekliliğini ortaya koymaktadır. Bu yüksek sayı , trafik kazalarını tetikleyebilecek veya yaralanma düzeyini artırabilecek risk faktörlerinin yeterince analiz edilmediğinin de bir göstergesi olarak alınabilir. Körfez İşbirliği Konseyi bölgesinde karayolu kazalarıyla ilişkili çeşitli risk faktörlerini araştırmak için bazı çalışmalar yapılmış olsa da, analizler için kullanılan verilerin eksikliği bu bölgedeki bir başka sorun olarak kendisini göstermektedir. Bu durum bu tez konusu ile ilgili araştırmaya olan ihtiyacı artırmaktadır. Bu nedenle, bu çalışma öncelikle trafik kazalarının ölüm ve yaralanma düzeyine etki eden potansiyel faktörleri araştırmayı amaçlamaktadır.

VS sınıflandırmak için öncelikle veri setinin önemli karakteristik özellikleri belirlenmiştir. İkinci olarak, sınıflandırma hedeflerine ulaşmak için literatürde sıklıkla kullanılan karar ağaçları sınıflandırma algoritması, veri setinden daha kesin bilgi ve ilişkileri keşfetmek için kullanılmıştır. Trafik kazalarının ciddiyetine yol açabilecek faktörleri incelemek için; yolun geometrisi, çarpışma türü, kazanın gerçekleşme zamanı, kazanın ana nedenleri, araç türü, yaralıların yaşları, yaralanma türleri ve cinsiyet faktörü dikkate alınmıştır Arama sonuçlarının Karayolu Genel Müdürlüğü, Emniyet Genel Müdürlüğü ile Sakarya Trafik Şube Dairesi de dahil olmak üzere ilgili tüm taraflara faydalı olacağı düşünülmektedir.

1.2. Problem Tanım

Trafik kazaları, dijital verilerin yoğun ve etkin olarak toplanıp kullanıldıkları alanlardan biridir. Bu nedenle, trafik kazası raporlarının analizi, merkezi hükümetlerin ve özel kuruluşların yol güvenliğine ilişkin planlarını etkileyebilir. Avrupa Yol Güvenliği Gözlemevi raporuna göre, Avrupa Birliği üye devletlerinde 2019 yılında

gerçekleşen trafik kazaları nedeniyle 26.100 ölüm meydana gelmiş iken 1,4 milyon kişi de yaralanmıştır [14]. Amerika Birleşik Devletleri'nde, trafik kazaları sonucu 36.461 kişi hayatını kaybetmiştir [15]. Orta Doğu bölgesi için, trafik kazalarından ölüm oranları 100.000 kişi için 22 kişi olarak gerçekleşmiştir. Bu rakama bakıldığında, Bahreyn dışındaki Konsey ülkelerine ait ölüm oranlarının, 10,6 ölüm oranına sahip olan Amerika Birleşik Devletleri'ndeki ölüm oranlarının oldukça üstünde olduğu görülmektedir. Aynı rapora göre Birleşik Arap Emirliklerindeki ölüm oranlarının 10,9 olarak gerçekleştiği belirlenmiştir. Veri madenciliği ulaşım mühendislerinin kaza analizleri ile ilgili çalışmalarında en çok kullandıkları araçlardan biri olarak önemini her gün artırmaktadır.

1.3. Araştırma İhtiyacı ve Önemi

Zamanımızda veri madenciliği teknikleri, çok özgün amaçlarla tamamen farklı alanlarda kullanılabilmektedir. Tüm ülkelerdeki polis merkezleri, suçu önlemek için suçları karakterize etmeye çalışarak, suçluların davranışları ve toplum için geçerli riskleri ve tehlikeleri sınıflandırırken veri madenciliğini kullandığı gibi, birçok çağrı merkezi, hizmet kalitesini iyileştirmek ve müşteri memnuniyetini artırmak için de bu tekniği kullanmaktadır. İnsan genomunun araştırılmasında, genleri ve işlevlerini keşfetmek için veri madenciliği teknikleri kullanılmıştır [16]. Diğer alanlarda başka örnekler den de bahsedilebilir. Bununla beraber veri madenciliğinin kullanımında ifade edilebilecek temel nokta, veri madenciliğinin karmaşık bir fenomeni daha iyi anlamak için karakterize etmeyi mümkün kılmasıdır. Problemi oluşturan faktör ve unsurlar arasındaki karşılıklı ilişki, etkileşim ve amaç fonksiyonuna olan tekil veya bileşik etkileri veri madenciliği sayesinde belirlenebilmektedir.

1.4. Çalışmanın Amacı ve Tez

Sakarya ilinde trafik kazalarından kaynaklanan çok sayıda ölüm ve yaralanma vakası, ilgili ölümleri ve yaralanmaları azaltmaya dönük bir çalışma yapılması gerekliliğini ortaya koymuştur. Bu yüksek sayı, trafik kazalarını tetikleyebilecek veya ölüm/yaralanma seviyesini artırabilecek risk faktörlerinin belirlenmesine dönük

eksikliğin bir göstergesi olarak ele alınabilir. Sakarya ilinde, trafik kazaları ile ilişkili çeşitli risk faktörlerinin karşılıklı ağırlıklı etkileşimlerini araştıran detaylı bir çalışma bulunmamaktadır. Bu çalışmaların yapılamamasındaki ana etkenlerden biri, kazaların kompleks yapılarının analizlerinde kullanılması gereken yeterli sayıda, çeşitlilikte ve derinlikte veri içeriğinin bulunmaması idi. Dijital olarak kaza verilerinin Sakarya Trafik Denetleme Şube Müdürlüğü tarafından kayıt altına alınması ve veri içeriklerinin zenginleştirilmesi, tez araştırma konusuna olan ihtiyacı da artırmaktadır. Dolayısı ile, bu çalışma ile Sakarya ili trafik kazaları ile ilgili veriler kullanılarak, gerçekleşen kazaların yapıları ve etkili faktörlerin kaza tipleri ve yapılarına olan tekil veya karşılıklı potansiyel etkileri ortaya konulmuştur. Kapsamlı karar ağaçları yöntemi bahsedilen etkinin belirlenmesinde ana yaklaşım yöntemi olarak kullanılmıştır. Yaralanma düzeyi sekiz farklı risk faktörü tarafından belirlenmiştir. Bayes ağ ve doğrusal destek vektör makine modelleri de en uygun modeli belirlemek için uygulanılmıştır. Trafik kazası yaralanmalarının seviyeleri: hafif, orta, şiddetli ve ölümcül yaralanma şeklinde dört gruba ayrılmıştır. Potansiyel risk faktörleri; yaralının cinsiyet ve yaşı, kaza türü, yol sınıfı, şerit sayısı, hız sınırı değerleri, emniyet kemeri kullanımı, kaza yeri, yaralı kişinin sürücü, yolcu veya yaya olması olarak belirlenmiştir. Bu çalışmanın sonuçları trafik kazası yaralanmalarına katkıda bulunan risk faktörlerinin potansiyel etkileri hakkında önemli yaklaşımsa ve stratejik değerlendirme olanakları vermektedir. Dolayısı ile belirlenen sonuçlar ışığında, etki eden risk faktörlerinin önceden belirlenmesi ve gerekli önlemlerin alınması ile ciddi veya ölümcül yaralanmalar hafifletilebilecektir. Bu anlamda karar vericiler yeni hukuki düzenlemeler yapabilir, ulaşım şebeke sistemlerinde var olan geometrik düzensizlikler iyileştirilebilir, işletme sistemlerinde yeni düzenlemeler yapılabilir.

1.5. Araştırma Metodu

Bu çalışmada benimsenen yöntem, karar ağaçları yöntemi, sınıflandırma için gerekli veri setinin önemli özellikleri ile uyumlu olduğu için tercih edilmiştir. Veri setinden net ve kesin bilgiler elde edilmesi içinde karar ağaçları oldukça etkin bir analiz metodudur. Metodun performans değerlendirmesi, sonuçların gösterdiği sınıflandırma kalitesine göre yapılacaktır.

Analiz süreci ile ilgili olarak aşağıdaki adımlar uygulanacaktır.

- Excel'de toplanan verilerin düzenlenmesi
- WEKA yazılımı kullanarak veri analizinin yapılması
- Karar ağaçlarının oluşturulması
- Sonuçların yorumlanması ve veri analizi değerlendirmesi

1.6. Tez İçeriği ve Organizasyonu

Tez içeriği ve sistematiği aşağıdaki gibi düzenlenmiştir.

1. BÖLÜM 1 (Giriş): Tezin bu ilk bölümü konunun önemi ve gerekliliği ile bu araştırmaya olan gereksinimi açıklamaktadır. Araştırma objektif ve amaçları da yine bu bölümde ele alınarak açıklanmıştır.

2. BÖLÜM 2 (Literatür Araştırması): İkinci bölümde, literatür araştırması çerçevesinde teze konu araştırma ile ilgili geçmişte ve günümüzde yapılan çalışmalara değinilerek, tezin teorik temelleri ortaya konulmuş ve çalışmanın güncel boyutu açıklanmıştır.

3. BÖLÜM 3 (Materyal Ve Yöntem): Üçüncü bölümde, araştırma yöntemi ayrıntılarıyla açıklanmıştır.

4. BÖLÜM 4 (Araştırma Bulguları): Dördüncü bölümde, araştırmanın sonuçları çıktılar halinde sunulmuş ve analiz edilmiştir.

5. BÖLÜM 5 (Sonuç ve Değerlendirme): Beşinci bölümde, analiz sonuçlarına dayanarak değerlendirmeler yapılarak gerekli öneriler ifade edilmiştir.

BÖLÜM 2. LİTERATÜR ARAŞTIRMASI

2.1. Veri Madenciliği Tarihçesi

Büyük miktarda veriden model oluşturmak ve analiz çalışması yapmak yeni bir olgu değildir. Model üretebilmek, veri toplama süreci ve kalitesi ile yakından ilgilidir. Çin'de, efsanevi İmparator Yao`nun ülkede üretilen mahsulleri MÖ 2238'de belirleme arzusu, Mısır'da firavun Amasis`in yaptırdığı nüfus sayımları veri toplama ve analiz çalışmalarına tarihte verilebilecek ilk örnekler arasındadır [17]. Bununla beraber, veriler arasındaki ortak özelliklerin arandığı veri analiz çalışmaları on yedinci yüzyılda başlamıştır. 1662'de John Graunt, Londra'daki ölümleri analiz ettiği ve hıyarcıklı vebanın salgınlarını tahmin etmeye çalıştığı "Mortalite Faturaları Üzerine Doğal ve Politik Gözlemler" adlı kitabını yayınladı. 1763'te Thomas Bayes, yalnızca bir deneyden kaynaklanan gözlemlerden olasılıkları değil, aynı zamanda bu olasılıklarla ilgili parametreleri de belirleyebileceğimizi gösterdi [17]. Binom dağılımının özel durumunda sunulan bu sonuç, Laplace tarafından bağımsız olarak genişletilerek Bayes teoreminin genel bir formülasyonuna öncülük etti [17]. Legendre, 1805'te bir veri kümesini matematiksel bir modelle karşılaştırmayı mümkün kılan en küçük kareler yöntemi üzerine bir makale yayınladı [17]. Bununla birlikte, pahalı, zahmetli ve zaman alıcı manuel hesaplamalar, bu yöntemlerin az sayıda basit ve aydınlatıcı durum dışında kullanılmasına izin vermemiştir.

Ronald Fisher, 1919'dan 1925'e kadar tıbbi istatistiksel çıkarım projesi için bir araç olarak varyans analizini geliştirdi [17]. 1950'ler göreceli olarak pahalı bilgisayarların ve toplu hesaplama tekniklerinin ortaya çıktığı dönem olmştur. Eşzamanlı olarak, segmentasyon, sınıflandırma, Perceptron adı verilen gelecekteki sinir ağlarının ilk versiyonu ve daha sonra genetik olarak adlandırılacak bazı kendi kendine gelişen algoritmalar gibi yöntemler ve teknikler ortaya çıkmıştır [17]. 1960'larda

araştırmacıların giderek daha hassas modellerden yararlanmasına ve analizler yapmasına olanak sağlayan karar ağaçları ve mobil merkez yöntemi veri madenciliğine kullanılmaya başlandı [18]. Fransa'da Jean-Paul Benzécri, 1962'de yazışma analizini geliştirdi [19].

1969'da, Myron Tribus'un otomatik hesaplama çerçevesinde Bayes yöntemlerini genelleştiren tanımları, kararları ve rasyonel kavramları ortaya çıktı [19] 1973'te "Rational Decisions in Uncertain" başlığı ile bu çalışma genişletilerek kullanıma sunuldu [20]. Çalışmanın önemli bir yönü, Cox-Jaynes teoreminden bahsedilmesidir. Böylece altta yatan bir sıklık çağrışımı olmaksızın, bir bilgi durumunun basit bir dijital çevirisi elde edilebilmektedir. Ayrıca, bu çalışma, olasılıkların Bayes kuralları çerçevesinde bir gözlemin sonuca katkısının ve etkisinin somut bir şekilde ölçülmesini mümkün kılmıştır.

Mikrobilgisayarların kademeli olarak gelişimi, Bayes yöntemlerinin ek maliyetler gerektirmeden genelleştirilmesini ve kullanımlarını kolaylaştırmıştır. Bu, Bayes kapsamlı analiz araştırmalarını teşvik etmiş, gözlemlere ait klasik istatistik sonuçlarına göre daha doğru ve etkin sonuçlar elde edildikçe de, kısa süreli zaman içeriklerinde rafine edilmiş bilginin üretilmesine olanak sağlamıştır.

2.2. Veri madenciliği İlkeleri ve Teorik Temelleri

VM veya Verilerde Bilgi Keşfi terimi genel olarak verilerin farklı perspektiflerden analizini ve veriler arasında ilişkileri kurarak veya kalıpları tespit ederek faydalı bilgilere dönüştürme eylemini ifade eder. Bu bilgiler daha sonra işletmeler tarafından geliri artırmak veya maliyetleri düşürmek için kullanılabilir. Bu tez konusu kapsamında ise kazaları ve kaza sonucu meydana gelen hasarları önlemek için daha etkin stratejiler oluşturmak amacıyla ilgili parametreler ve unsurlar arasındaki ilişkilerin ortaya konulması amacı ile veri madenciliği kullanılmıştır.

Veri Madenciliği yazılımı, veri analizi için kullanılan analitik araçlardan biridir. Kullanıcıların verileri farklı açılardan analiz etmelerine, kategorilere ayırmalarına ve

belirlenen ilişkileri özetlemelerine olanak tanır. Teknik olarak Veri Madenciliği, çok sayıda ilişkisel veri tabanı arasındaki korelasyonları veya modelleri bulmaya izin veren bir süreç olarak ifade edilebilir. Bu anlamda Veri Madenciliği, verileri bölümlere ayırmak ve gelecekteki olasılıkları değerlendirmek için karmaşık ve kompleks algoritmalara dayanır.

2.2.1. Doğal bir teknolojik dönüşüm

Veri Madenciliği terimi nispeten yeni olmakla beraber, analizlerde kullanılan teknolojiler göreceli olarak çok yeni değildir. Şirketler, büyük hacimli verileri işlemek ve pazar araştırma raporlarını analiz etmek için yıllardır güçlü bilgisayarlar kullanmaktadırlar. İstatistiksel yazılım alanlarındaki sürekli gelişen yenilikler, analizlerin doğruluğunu büyük ölçüde artırmakta ve maliyeti düşürmektedir.

2.2.2. Veri Madenciliğinde veri, bilgi ve bilin

1. Veri

Veriler, bir bilgisayar tarafından işlenebilen sayılar veya metinlerdir. Günümüzde şirketler, farklı biçimlerde, büyük miktarlarda veri biriktirmektedirler. Veriler içerikleri itibari ile genel anlamda aşağıdaki gibi kategorize edilebilir.

- Operasyonel veya işlemse veriler (satışlar, maliyetler, envanter, makbuzlar veya muhasebe verileri gibi.)
- İlişkisel olmayan veriler (Endüstriyel satışlar, tahmin verileri, makroekonomik veriler gibi.)
- Meta veriler.

2. Bilgi

Tüm bu veriler arasındaki ilişkiler, kurulan modeller ve algoritmalar sayesinde elde edilebilmektedir. Örneğin, satış noktası işlem verilerini analiz etmek, hangi ürünlerin

satıldığı ve bu satışların ne zaman gerçekleştiği hakkında bilgi değerlendirme sürecinin sonuçlarının elde edilmesini sağlamaktadır.

3. Sonuç

Geçmiş yapılar veya gelecekteki eğilimler hakkındaki ilişkiler bilgiye dayalı köprüler sayesinde kurulabilir. Örneğin, bir süpermarketin perakende satışları hakkındaki bilgiler, alıcıların davranışları hakkında bilgi edinmek amaçlı promosyon çabalarının bir parçası olarak analiz edilebilir. Böylece, bir üretici veya perakendeci, veri madenciliği analiz sonuçlarını kullanarak hangi ürünlerin tanıtılması gerektiğini belirleyebilir.

2.2.3. Veri madenciliği yöntemleri

Beş çeşit Veri Madenciliği vardır:

- İlişkilendirme: bir olayın başka bir olayla bağlantılı olduğu kalıpların, yapıların belirlenmesi.
- Sıra analizi: bir olayın daha sonraki bir olaya yol açtığı etki kalıplarının belirlenmesi.
- Sınıflandırma: verilerin düzenlenme ve gruplanma şekillerinin ortaya konulması.
- Kümeleme: bilinmeyen ilişkilere ait kümelerin bulunması ve görsel olarak belgelenmesi.
- Tahmin: Gelecekle ilgili makul öngörülerin ve değerlendirmelerin yapılması.

Dolayısı ile veri madenciliği tahmine dayalı analitik bir süreç olarak algılanmaktadır.

2.2.4. Veri madenciliği nasıl çalışır

Veri Madenciliği, birbirinden ayrı gelişen bilgisayar teknolojileri ile işlemsel ve analitik sistemler arasındaki bağlantıyı sağlamaktadır. Veri madenciliği yazılımı,

kullanıcı sorgularına dayalı olarak depolanan işlem verilerinin ilişkilerini ve modellerini analiz etmektedir. İstatistiksel, makine öğrenimi ve sinir ağları gibi çeşitli analitik yazılım türleri mevcuttur. Modellerin temelde çalışma süreçleri olarak dört tür aşama söz konusudur:

- Sınıflama: Depolanan verilerin, belirlenmiş gruplara yerleştirmesi sürecidir. Örneğin, bir restoran zinciri, müşteri ziyaretlerinin ne zaman gerçekleştiğini ve normal siparişlerinin ne olduğunu belirlemek için müşteri satın alma verilerini araştırabilir. Bu bilgiler, günlük menüler sunarak müşteri sayısını artırmak da kullanılabilir.
- Kümeleme: Veriler, mantıksal ilişkilere veya müşteri tercihlerine göre birlikte gruplandırılır. Örneğin, pazar portföylerini veya müşteri yakınlıklarını belirlemek için veriler alt sınıflamada kümelenebilir.
- İlişkilendirme: Kümelenen veriler gerek kendi içlerinde gerek diğer küme elemanları ile ilişkilendirilerek, parametrik etkileşim tanımlamaları yapılabilir.
- Sıralı Modelleme: Ana yapıdaki davranış kalıplarını ve eğilimleri tahmin etmek üzere farklı parametrelerin ortak etkileşimlerinin doğurduğu neticelere ait yapılar, sıralı modellemeler ile elde edilebilir. Örneğin, bir dış mekân ekipman satıcısı, müşterinin uyku tulumu ve yürüyüş ayakkabısı satın alımlarına dayanarak bir sırt çantasının satın alınma olasılığını tahmin edebilir.

Aşağıdaki şekil karmaşık veri yapılarından düzenli sistemlerin ve sonuçların elde edilmesine ait görsel olarak değerlendirilebilir.

Şekil 2.1. Veri madenciliği-Kaostan düzene.

Veri madenciliği analizlerinde aşağıdaki beş ana aşama temel işlem adımları olarak görülebilir.

- İşlem verilerinin ayıklanması, dönüştürülmesi ve Veri Ambarı sistemine yüklenmesi.
- Verilerin çok boyutlu bir veri tabanı sisteminde depolanması ve yönetimi.
- Analistlerin ve BT uzmanları için veri erişiminin sağlanması.
- Uygulama yazılımlarının kullanılarak verilerin analiz edilmesi.
- Verilerin grafik veya tablo gibi kullanışlı bir biçimde sunulması.

Veri madenciliği analizlerinde farklı uygulama yöntemleri ve algoritmalar kullanılmaktadır.

- Yapay Sinir Ağları: Eğitim yoluyla öğrenen ve biyolojik sinir ağlarına benzeyen doğrusal olmayan tahmin modelleri.
- Genetik Algoritmalar: Doğal tekâmül kavramlarına dayanan bir tasarımla genetik kombinasyon, mutasyon ve doğal seleksiyon gibi süreçleri kullanan optimizasyon teknikleri.
- Karar Ağaçları: Bu ağaç şeklindeki yapılar, karar dizilerini temsil etmektedir. Karar dizileri, veri kümesini sınıflandırmak için oluşturulan kurallardan oluşur. Özel karar ağacı yöntemleri arasında Sınıflandırma ve Regresyon Ağaçları (CART) ve Chi Kare Otomatik Etkileşim Algılama (CHAID) bulunur. Bu iki yöntem, bir veri setinin sınıflandırılması için kullanılır. Hangi verilerin nasıl bir sonuç üreteceğini tahmin etmek için yeni bir veri kümesine uygulanabilecek bir dizi kural sağlarlar. CART, iki yönlü bir bölüm oluşturarak bir veri kümesini bölümlere ayırırken, CHAID çok yönlü sonuçlar oluşturmak için seti ki-kare testlerini kullanarak bölümlere ayırır. Genel olarak, CART, CHAID'den daha az veri hazırlığı gerektirmektedir.
- En Yakın Komşu Yöntemi: Bu teknik, bir veri kümesindeki her veriyi, özellikleri belli bir veri kümesine olan benzerliği dolayısı ile mevcut farklı kombinasyonlara dayalı olarak sınıflandırır.

- Kural indüksiyonu: İstatistiksel anlamlılığa dayalı olarak verilerden "eğer-ise" kurallarının çıkarılması esasına dayalı olarak işlemektedir.
- Veri görselleştirme: Çok boyutlu verilerdeki karmaşık ilişkilerin görsel yorumlamalarının yapılması için farklı grafik araçlar kullanılır.

Yukarıdaki açıklamalar ışığında genel süreçte ilk olarak, veriler toplanır ve veri ambarlarına yüklenir. Bu şekilde veriler, fiziksel sunucularda veya Bulutta depolanmış olunur. Bu verilere erişen analistler ve uzmanlar öncelikli olarak bu verileri nasıl düzenlemek istediklerini belirlerler. Ardından, uygulama yazılımı, verileri kullanıcı sonuçlarına göre sıralamayı mümkün kılar. Son olarak, son kullanıcı verileri grafik veya tablo gibi paylaşması kolay bir formatta sunularak analiz süreci tamamlanmış olunur.

Büyük veri madenciliği temelde bakıldığında şu 3 ana özelliğe sahiptir

- Otomatik ilişki keşfi

Veri Madenciliği, modellerin geliştirilmesine dayanır. Bir model, bir dizi veriye göre hareket etmek için bir algoritma kullanır. Otomatik keşif kavramı, veri madenciliği modellerinin yürütülmesine atıfta bulunur. Veri Madenciliği modelleri, üzerine inşa edildikleri veriler arasındaki karmaşık ilişkileri ortaya çıkarmak için kullanılırsa da çoğu model türü yeni verilere genelleştirilebilir.

- Olası sonuçları tahmin etmek

Veri Madenciliğinin birçok biçimi öngörücüdür. Örneğin, bir model eğitim ve diğer demografik faktörlere dayalı olarak bir sonucu tahmin edebilir. Tahminlerin ilişkili bir olasılığı vardır. Tahmine dayalı veri madenciliğinin bazı biçimleri, bir sonuç elde etmenin koşulları olan kurallar üretir. Örneğin, bir kural, belirli bir kaza tipinin sürücü yaşına veya cinsiyetine bağlı olarak gerçekleşme olasılığının yüksek olabileceğini belirtebilir.

- Eyleme dönüştürülebilir bilgilerin oluşturulması

Veri Madenciliği, büyük hacimli verilerden, yararlanılabilir bilgilerin çıkarılmasını mümkün kılar. Örneğin, bir şehir planlamacısı, düşük gelirli haneler için bir plan geliştirmek üzere demografiye dayalı geliri tahmin etmek için bir model kullanabilir. Bir araba kiralama ajansı, yüksek değerli müşterileri hedefleyen bir promosyon oluşturmak için tüketici seğmenlerini tanımlamak için bir model kullanabilir.

2.2.5. Teknolojik altyapı

Günümüzde Veri Madenciliği uygulamaları ana bilgisayar, sunucu veya PC ölçeğinde her boyutta yapılabilmektedir. Sistem fiyatları, en küçük uygulamalar için birkaç bin dolardan, büyük uygulamalar için terabayt başına 1 milyon dolara kadar değişebilmektedir. İş amaçlı yapılan uygulamalar genellikle 10 gigabayttan 11 terabaytın üzerine kadar değişmektedir. NCR, Data Center in Delhi, 100 terabayttan fazla uygulama sunma kapasitesine sahiptir. Veri madenciliğinde iki ana teknolojik faktörden söz edilebilir:

- Veritabanı boyutu: İşlenecek ve sürdürülecek ne kadar çok veri olursa, o kadar güçlü bir sisteme ihtiyaç duyulacaktır.
- Taleplerin karmaşıklığı: Talepler ne kadar karmaşık ve çoksa, o kadar güçlü bir sistem gerekir.

İlişkisel Veritabanı depolama ve yönetim teknolojileri, 50 gigabaytın altındaki birçok veri madenciliği uygulaması için yeterlidir. Ancak, daha büyük uygulamaları desteklemek için bu altyapının büyük ölçüde artırılması gerekir. Bazı analistler, sorgu performansını artırmak için daha büyük dizin oluşturma yeteneği olan sistemler geliştirdiler. Diğerleri, sorgu işleme süresini iyileştirmek için Massiely Parallel Processors (MPP) gibi yeni donanım mimarilerini kullanmayı tercih etmektedirler. Örneğin, NCR'nin MPP sistemleri, en iyi süper bilgisayarlardan daha yüksek performans seviyeleri elde etmek için yüzlerce Pentium işlemciyi birbirine bağlayabilmektedir.

Şekil 2.2. İşletim bilgilerinin oluşturulması

2.2.6. Veri madenciliği Yazılımı

Veri Madenciliği yazılımı, veriler arasındaki ilişkileri analiz eder ve kullanıcı isteklerine göre ilişkilere ait kalıpları ve ortak çıkarımları tanımlar ve bilgi sınıfları oluşturur. Örneğin, kazaların ne zaman meydana geldiği ve kaza tipleri ile ilgili olarak toplanan veriler ışığında, günün, haftanın, ayın hatta yıllın hangi dönemlerinde hangi tip kazaların hangi boyutta gerçekleştiği ve hasarın ölçeği ile ilgili sınıflar oluşturmak mümkün olabilmektedir.

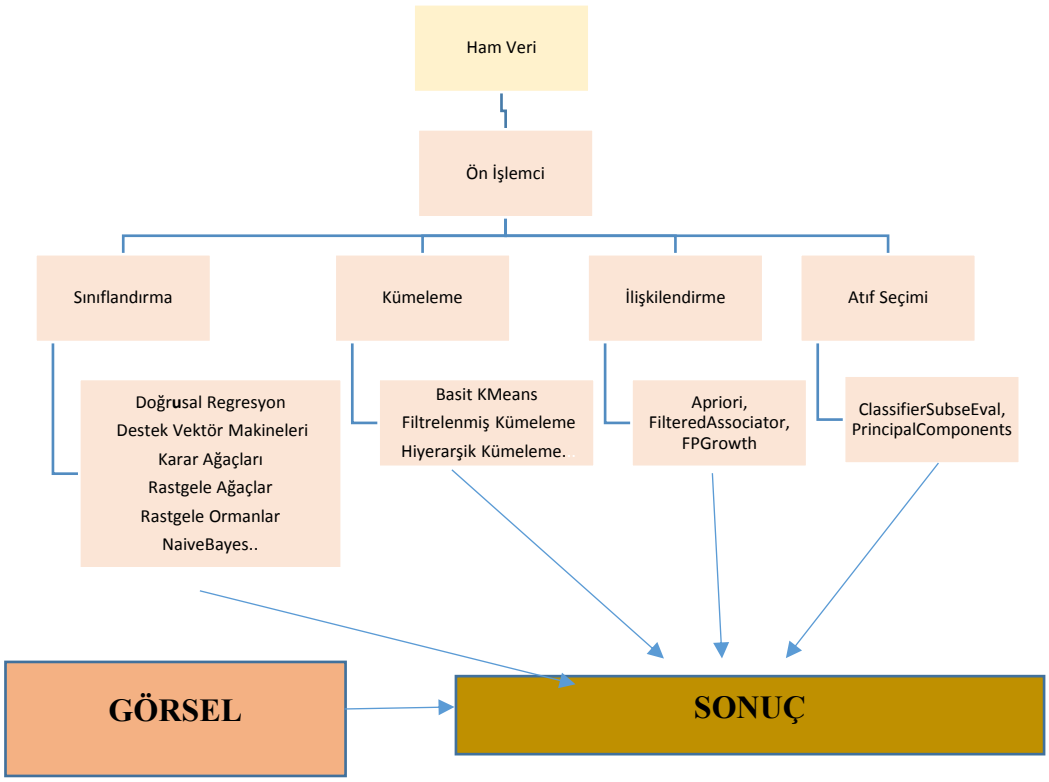
Diğer durumlarda, veri madencileri mantıksal ilişkilere dayalı bilgi kümeleri bulurlar veya kullanıcı davranışı hakkında sonuçlar çıkarmak için ilişkilendirmeleri ve sıralı kalıpları ararlar. Orange, Weka, RapidMiner veya Tanagra, web'de bulunan açık kaynaklı veri madenciliği yazılım araçlarından bazılarıdır. Veri Madenciliği için profesyonel lisanslar da mevcuttur. Bunların en popülerleri olarak IBM tarafından geliştirilen SPSS, SAS'tan Enterprise Miner veya Redmond firmasından Microsoft Analysis Services zikredilebilir.

2.3. WEKA

WEKA, Yeni Zelanda'daki Waikato Üniversitesi tarafından geliştirilen bir veri madenciliği sistemidir. Veri madenciliği algoritmalarını uygulayan WEKA, son teknoloji ürünü bir yazılımdır. Makine öğrenimi (ML) teknikleri ve bunların gerçek dünya veri madenciliği problemlerine uygulanması içeriği ile oldukça yaygın bir

kullanıma sahiptir. WEKA; veri ön işleme, sınıflandırma, regresyon, kümeleme, ilişkilendirme kuralları, ayrıca elde edilen sonuçlar için görselleştirme araçları içermektedir.

WEKA'nın genel yapısı aşağıdaki şekilde sunulmuştur.



Şekil 2.3. WEKA genel yapısı

2.3.1. Weka explorer

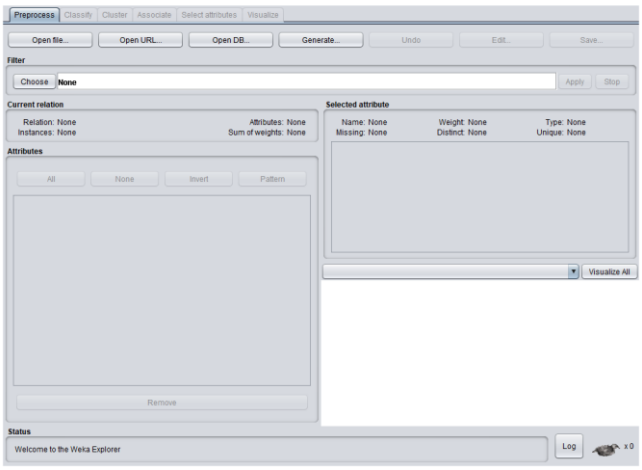
WEKA GUI Chooser uygulaması başladığında ve aşağıdaki ekranı görülmektedir.

Şekil 2.4. WEKA başlangıç ara yüzü görüntüsü

GUI Chooser uygulaması, burada listelendiği gibi beş farklı türde uygulamanın çalıştırılmasına izin vermektedir.

- Keşfedici
- Deneyci
- Bilgi Akışı
- Tezgah
- Basit CLI

Applications selector'da Explorer butonuna tıkladığımızda aşağıdaki ekran açılır



Şekil 2.5. Keşfedici (Explorer) ara yüzü

En üstte listelendiği gibi Explorer ara yüzünde birkaç sekme vardır

- Ön işlem
- Sınıflandırma
- Kümeleme
- İlişkilendirme
- Öznitelik Seçimi
- Görselleştirme

Bu sekmelerin altında, önceden uygulanmış birkaç makine öğrenimi algoritması vardır. Aşağıda bunların her birine ayrıntılı olarak değinilmiştir.

1. Ön İşlem Sekmesi

Başlangıçta, gezgin açıldığında yalnızca Ön İşlem sekmesi etkinleştirilir. Makine öğreniminin ilk adımı, verileri önceden işlemektir. Böylece, Önişleme seçeneğinde, veri dosyası seçilmiş, işlenmiş ve çeşitli makine öğrenme algoritmaları uygulamaya uygun hale getirilmiş olur.

2. Sekmeyi Sınıflandır

Sınıflandır sekmesi, verilerin sınıflandırılması için kullanıcıya birkaç makine öğrenimi algoritması sağlar. Bu amaç için Doğrusal Regresyon, Lojistik Regresyon, Destek Vektör Makineleri, Karar Ağaçları, Random Tree, Random Forest, Naive Bayes gibi algoritmalar uygulanabilmektedir. Liste çok kapsamlıdır ve hem denetimli hem de denetimsiz makine öğrenimi algoritmaları kullanılabilmektedir.

3. Küme Sekmesi

Küme sekmesinin altında, Simple K-Means, Filtered Clusterer, Hierarchical Clusterer vb. Gibi birkaç kümeleme algoritmaları sunulmaktadır.

4. İlişkilendirme Sekmesi

İlişkilendirme (Associate) sekmesinin altında, Apriori, Filtered Associator ve FP Growth kullanılabilmektedir.

5. Özellikler Sekmesi

Nitelikleri Seçme, Classifier Subset Eval, Prinicipal Components, vb. gibi çeşitli algoritmalara dayalı seçimlerin yapılabilmesine olanak tanır.

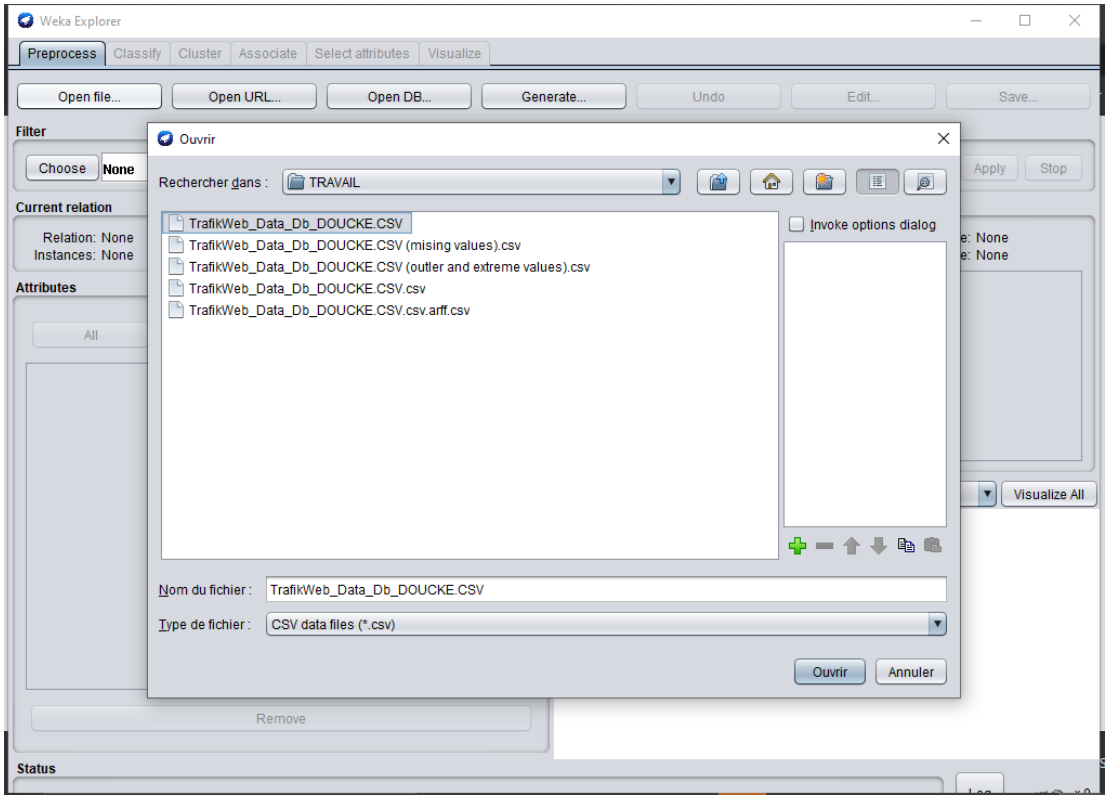
6. Sekmeyi Görselleştir

Görselleştir seçeneği, işlenmiş verilerin analiz için görselleştirilmesini sağlamaktadır.

Açıklamalardan görülebileceği gibi, WEKA, makine öğrenimi uygulamalarını test etmek ve oluşturmak için birkaç kullanıma hazır algoritma sağlamaktadır. WEKA'nın etkili bir şekilde kullanılması, bu algoritmaların nasıl çalıştıkları, hangi koşullar altında hangisinin seçileceği, işlenmiş çıktılarda neye dikkat edileceği gibi alanlarda yeterli bir bilgiye sahip olunmasını gerektirmektedir. Dolayısı ile, WEKA uygulamalarını oluştururken ve etkili bir kullanım için makine öğrenimi literatüründe sağlam bir temele ihtiyaç duyulmaktadır.

2.3.2. Yerel dosya sisteminden veri yükleme

Aşağıdaki görsel bir dosyanın Weka'da yerel olarak nasıl içe aktarılacağını göstermektedir:

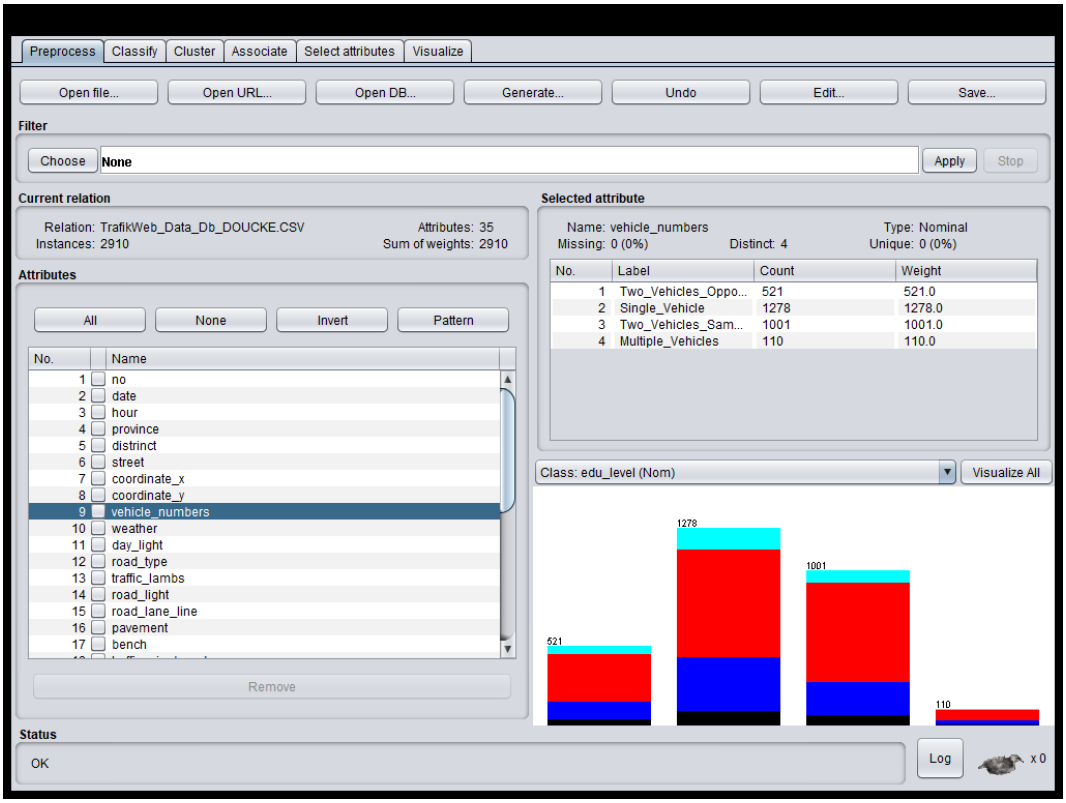


Şekil 2.6. Yerel Dosyanın WEKA ya aktarımı

Dosya WEKA`ya aktarılmadan önce dosyanın CSV formatına çevrilmesi gerekir. Dosya CSV formatına dönüştürüldükten sonra dosya uygun bir klasöre kaydedilir ve ardından aşağıdaki işlem adımları uygulanır:

- Weka yazılımını açın
- Keşfetme tıklayın
- Bir dosya açın
- Dosyayı seçin
- Dosya türünü seçin
- Dosyayı CSV formatında seçin, ardından aç'a tıklayın

Açılan dosya WEKA analiz sürecine bu aşamadan sonra dahil edilebilir.



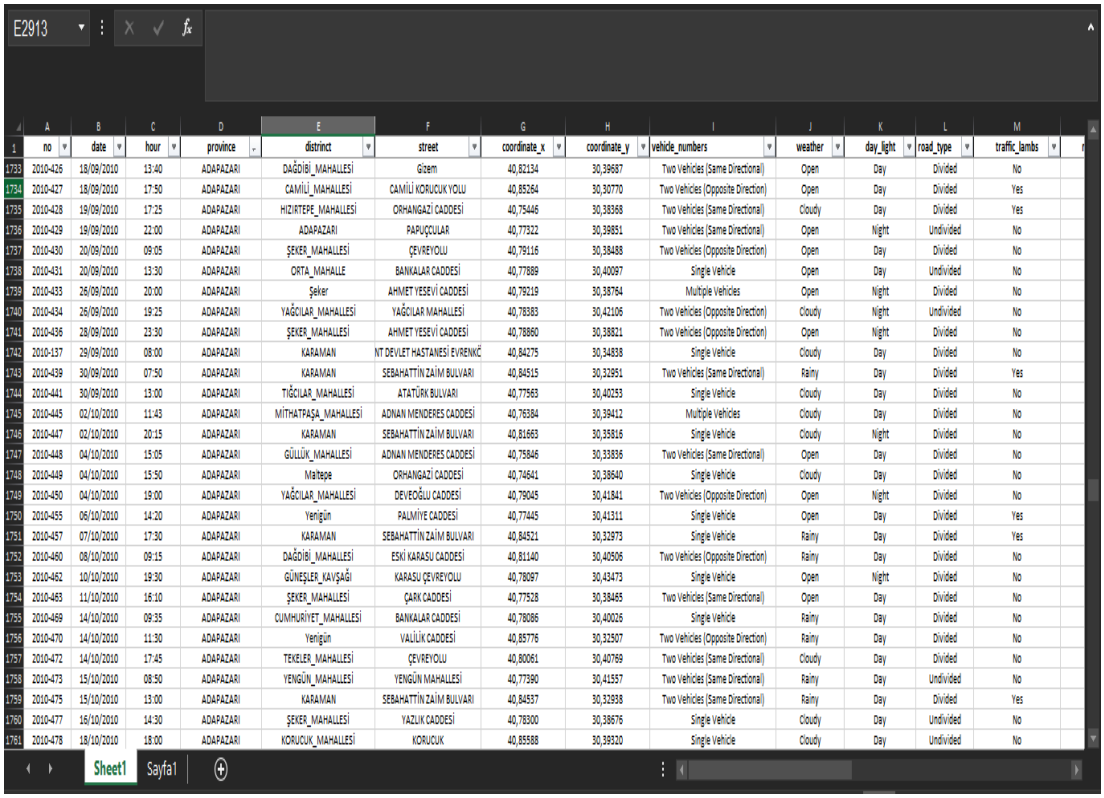
Şekil 2.7. her bir özellik ve içeriği

Şekil 2.7. veri seti içeriğinde bulunan her bir parametreye ait özelliklerin içeriğiyle birlikte görselleştirilmesini sağlar. Veri ön işleme sürecinden sonra elde edilecek. karar ağacı geliştirme süreci için bu görselleştirme çok önemli bir adımdır.

BÖLÜM 3. MATERYAL VE YÖNTEM

3.1. Materyal

Bu çalışmada Sakarya İl Emniyet Müdürlüğü'nden elde edilen ve Sakarya ilinde meydana gelen trafik kazaları veri madenciliği yöntemleri ile analiz edilmiştir. Bu araştırmanın veri setinde, meydana gelen kazalarla ile ilgili 139 özellik sınıfı bulunmaktadır.



	A	B	C	D	E	F	G	H	I	J	K	L	M
1	no	date	hour	province	distrinct	street	coordinate_x	coordinate_y	vehicle_numbers	weather	day_light	road_type	traffic_lambs
1733	2010-426	18/09/2010	13:40	ADAPAZARI	DAĞDİBİ_MAHALLESİ	Gizem	40,82134	30,39687	Two Vehicles (Same Directional)	Open	Day	Divided	No
1734	2010-427	18/09/2010	17:50	ADAPAZARI	CAMİLİ_MAHALLESİ	CAMİLİ KORUCUK YOLU	40,85264	30,30770	Two Vehicles (Opposite Direction)	Open	Day	Divided	Yes
1735	2010-428	19/09/2010	17:25	ADAPAZARI	HIZIRTEPE_MAHALLESİ	ORHANGAZİ CADDESİ	40,75446	30,38368	Two Vehicles (Same Directional)	Cloudy	Day	Divided	Yes
1736	2010-429	19/09/2010	22:00	ADAPAZARI	ADAPAZARI	PAPUÇCULAR	40,77322	30,39851	Two Vehicles (Same Directional)	Open	Night	Undivided	No
1737	2010-430	20/09/2010	09:05	ADAPAZARI	ŞEKER_MAHALLESİ	ÇEVREYOLU	40,79116	30,38488	Two Vehicles (Opposite Direction)	Open	Day	Divided	No
1738	2010-431	20/09/2010	13:30	ADAPAZARI	ORTA_MAHALLE	BANKALAR CADDESİ	40,77889	30,40097	Single Vehicle	Open	Day	Undivided	No
1739	2010-433	26/09/2010	20:00	ADAPAZARI	Şeker	AHMET YESEVİ CADDESİ	40,79219	30,38764	Multiple Vehicles	Open	Night	Divided	No
1740	2010-434	26/09/2010	19:25	ADAPAZARI	YAĞCILAR_MAHALLESİ	YAĞCILAR MAHALLESİ	40,78383	30,42106	Two Vehicles (Opposite Direction)	Cloudy	Night	Undivided	No
1741	2010-436	28/09/2010	23:30	ADAPAZARI	ŞEKER_MAHALLESİ	AHMET YESEVİ CADDESİ	40,78860	30,38821	Two Vehicles (Opposite Direction)	Open	Night	Divided	No
1742	2010-137	29/09/2010	08:00	ADAPAZARI	KARAMAN	NT DEVLET HASTANESİ EVRENKÖ	40,84275	30,34838	Single Vehicle	Cloudy	Day	Divided	No
1743	2010-439	30/09/2010	07:50	ADAPAZARI	KARAMAN	SEBAHATTİN ZAİM BULVARI	40,84515	30,32951	Two Vehicles (Same Directional)	Rainy	Day	Divided	Yes
1744	2010-441	30/09/2010	13:00	ADAPAZARI	TİĞCILAR_MAHALLESİ	ATATÜRK BULVARI	40,77563	30,40253	Single Vehicle	Cloudy	Day	Divided	No
1745	2010-445	02/10/2010	11:43	ADAPAZARI	MİTHATPAŞA_MAHALLESİ	ADNAN MENDERES CADDESİ	40,76384	30,39412	Multiple Vehicles	Cloudy	Day	Divided	No
1746	2010-447	02/10/2010	20:15	ADAPAZARI	KARAMAN	SEBAHATTİN ZAİM BULVARI	40,81663	30,35816	Single Vehicle	Cloudy	Night	Divided	No
1747	2010-448	04/10/2010	15:05	ADAPAZARI	GÜLLÜK_MAHALLESİ	ADNAN MENDERES CADDESİ	40,75846	30,33836	Two Vehicles (Same Directional)	Open	Day	Divided	No
1748	2010-449	04/10/2010	15:50	ADAPAZARI	Maltepe	ORHANGAZİ CADDESİ	40,74641	30,38640	Single Vehicle	Cloudy	Day	Divided	No
1749	2010-450	04/10/2010	19:00	ADAPAZARI	YAĞCILAR_MAHALLESİ	DEVEOĞLU CADDESİ	40,79045	30,41841	Two Vehicles (Opposite Direction)	Open	Night	Divided	No
1750	2010-455	06/10/2010	14:20	ADAPAZARI	Yenigün	PALMİYE CADDESİ	40,77445	30,41311	Single Vehicle	Open	Day	Divided	Yes
1751	2010-457	07/10/2010	17:30	ADAPAZARI	KARAMAN	SEBAHATTİN ZAİM BULVARI	40,84521	30,32973	Single Vehicle	Rainy	Day	Divided	Yes
1752	2010-460	08/10/2010	09:15	ADAPAZARI	DAĞDİBİ_MAHALLESİ	ESKİ KARASU CADDESİ	40,81140	30,40506	Two Vehicles (Opposite Direction)	Rainy	Day	Divided	No
1753	2010-462	10/10/2010	19:30	ADAPAZARI	GÜNEŞLER_KAVŞAĞI	KARASU ÇEVREYOLU	40,78097	30,43473	Single Vehicle	Open	Night	Divided	No
1754	2010-463	11/10/2010	16:10	ADAPAZARI	ŞEKER_MAHALLESİ	ÇARK CADDESİ	40,77528	30,38465	Two Vehicles (Same Directional)	Open	Day	Divided	No
1755	2010-469	14/10/2010	09:35	ADAPAZARI	CUMHURİYET_MAHALLESİ	BANKALAR CADDESİ	40,78086	30,40026	Single Vehicle	Rainy	Day	Divided	No
1756	2010-470	14/10/2010	11:30	ADAPAZARI	Yenigün	VALİLİK CADDESİ	40,85776	30,32507	Two Vehicles (Opposite Direction)	Rainy	Day	Divided	No
1757	2010-472	14/10/2010	17:45	ADAPAZARI	TEKELER_MAHALLESİ	ÇEVREYOLU	40,80061	30,40769	Two Vehicles (Same Directional)	Cloudy	Day	Divided	No
1758	2010-473	15/10/2010	08:50	ADAPAZARI	YENGÜN_MAHALLESİ	YENGÜN MAHALLESİ	40,77390	30,41557	Two Vehicles (Same Directional)	Rainy	Day	Undivided	No
1759	2010-475	15/10/2010	13:00	ADAPAZARI	KARAMAN	SEBAHATTİN ZAİM BULVARI	40,84537	30,32938	Two Vehicles (Same Directional)	Rainy	Day	Divided	Yes
1760	2010-477	16/10/2010	14:30	ADAPAZARI	ŞEKER_MAHALLESİ	YAZLIK CADDESİ	40,78300	30,38676	Single Vehicle	Cloudy	Day	Undivided	No
1761	2010-478	18/10/2010	18:00	ADAPAZARI	KORUCUK_MAHALLESİ	KORUCUK	40,85588	30,39320	Single Vehicle	Cloudy	Day	Undivided	No

Şekil 3.1. Sakarya kaza verileri formatı

Yukarıdaki Şekil 3.1., WEKA yazılımı ile analiz edeceğimiz ve sınıflandırmasını yapacağımız verilerin bazı temel içeriklerini Excel formatında sunmaktadır.

3.2. Veri Toplama Yöntemi

Bu çalışmada, Sakarya Emniyet Müdürlüğü trafik müdürlüklerinden elde edilen kazalara ilişkin veriler analiz edilmiştir. Veriler, 2006 ve 2010 yılları arasındaki 2.911 örneği içermektedir. Bu tez içeriğinde incelenen trafik kazası verileri, tümüyle Sakarya şehrinde kayıtlı kazalara ve bunların mevcut değişkenlerine bağlıdır.

Elde edilen kaza verileri, kaza sayılarını, yaralanma seviyelerini, kazaların nedenlerini ve araç tipleri gibi öznitelikleri (parametreleri) karakterize etmektedir. 1 Ocak 2006 ile 27 Aralık 2010 tarihleri arasında emniyet mensupları tarafından derlenen veriler, rapor edildikleri şekliyle ham veri halinde olduğundan, tez analiz çalışması için ihtiyaç duyulan veri formatından oldukça farklıydılar. Bu nedenle, ölümlü ve ölümlü olmayan trafik kazalarının derecesine modellemek için önemli değişkenler belirlenmiş ve kategorize edilmek üzere kodlanmıştır. Bu çalışmada yol geometrisi, çarpışma türü, kaza zamanı, kaza nedenleri, araç türü, kazazedenin yaşı, hava yastığı durumu, cinsiyet, hava durumu, gün ışığı, yol türü, trafik sıkışıklıkları, yol şerit çizgisi, kaldırım, trafik tabelası, yol yönü, yüzey, araç tipi, sürücü eğitim seviyesi, ehliyet sahipliği, alkol durumu gibi 18 bağımsız değişken (öznitelik-parametre) dikkate alınmış ve incelenmiştir.

3.2.1. Verilerin toplaması

Tez içeriğinde analiz amacı ile kullanılan veriler, 2010 – 2016 yılları arasında il genelinde gerçekleşen trafik kazalarında polis tarafından Trafik Kaza Bilgi Sistemi kapsamında toplanmış ve Excel'de kayıt altına alınmıştır.

Şekil 3.2. Sakarya Trafik kazası

3.3. Yöntem

Bu çalışmada, veri setinin önemli özelliklerini sınıflandırmak için WEKA yazılımı seçilmiştir. Veri setinden daha kesin, net bilgiler ve ilişkiler keşfetmek ve sınıflandırma hedeflerine ulaşmak için de literatürde sıklıkla kullanılan bir sınıflandırma algoritması olan karar ağaçları kullanılmıştır. Bu algoritmaların performans değerlendirmesi, sonuçların gösterdiği sınıflandırma kalitesine göre yapılmıştır.

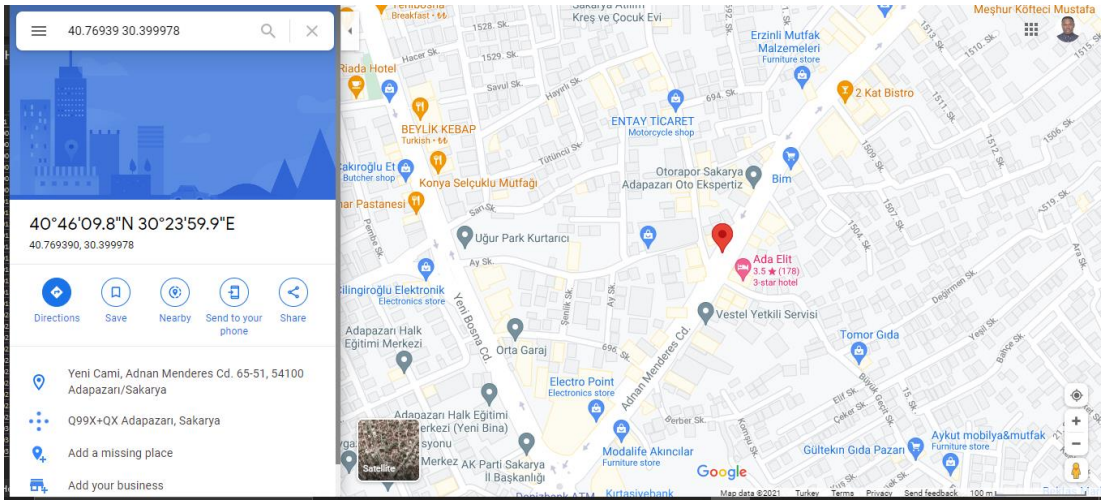
3.3.1. Veri Ön İşleme

Ham verilerin sağlıklı analizi için, veriler içinde bulunan hatalı ve anlamsız verilerin veri kümesinden çıkarılması veya düzenlenmesi son derece önemlidir. B amaçla anlamsız, hatalı veya tutarsız veriler titiz bir ön inceleme ve çalışma ile veri setinden çıkarılmıştır. Bu süreçleri de içeren WEKA analizi aşağıda belirtine ön işleme süreci ve adımları çerçevesinde yapılmıştır.

- Excel'de verilerin düzenlenmesi
- WEKA üzerinde maskeleme verilerinin işlenmesi
- Aykırı Değerlerin ve Uç Değerlerin Weka'da İşlenmesi
- Weka'da veri analizi
- Weka'da karar ağacı işlemleri

3.3.1.1. Excel verilerinin düzenlenmesi

Bu adımda, verilerde esik olan mahalle veya ilçe kodları ile adlarına ait veriler, excel dosyasında bulunan koordinat bilgileri Googla Map uygulamasına yüklenerek elde edilmiştir.



Şekil 3.3. Google haritalarda mahalle ve semt araması

Google Maps üzerinde mahalle ve sokakları X ve Y koordinatları üzerinden tanımladıktan sonra WEKA üzerinde veri ön işlemesine geçilmiştir.

3.3.1.2. Weka'da veri önişleme

Pencerenin en üstünde, başlık çubuğunun hemen altında bir dizi sekme bulunur. Sadece ilk 'Önişleme' sekmesi, açık veri kümesi olmadığı için etkindir. İlk üç Önişleme bölümünün üst kısmındaki 4 düğme, WEKA'ya veri yüklenmesini sağlamaktadır. Veri, yukarıda açıklandığı gibi mevcut bir dosyadan çeşitli biçimlerde ARFF, CSV,

C4.5 formatında içe aktarılabilmektedir. Ayrıca URL veya bir SQL veri tabanından JDBC kullanılarak, harici bir dosyadan da veriler okunabilmektedir.

3.3.1.2.1. Weka'da eksik verilerin tespit edilmesi ve düzenlenmesi

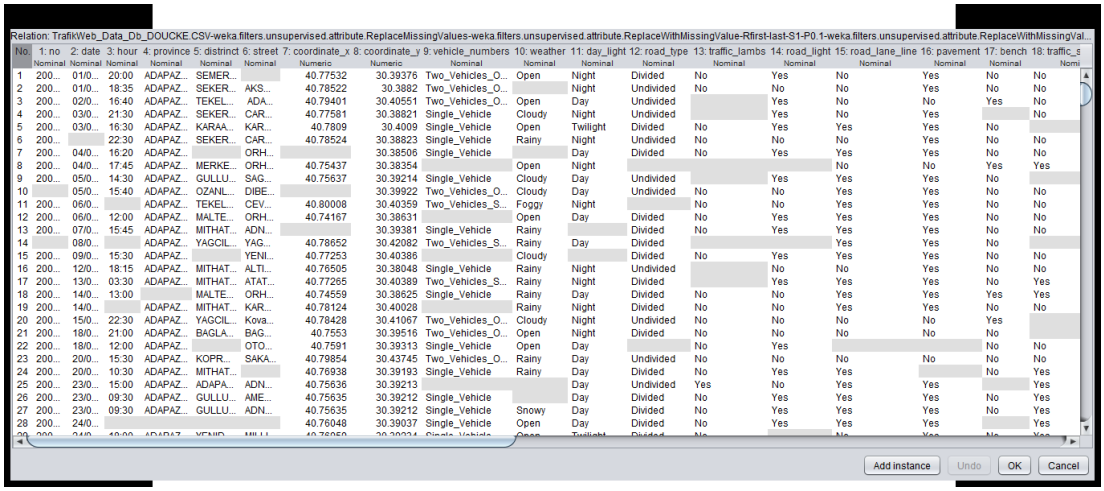
- Eksik veriler nasıl tespit edilir

Weka'da eksik verileri tespit etmek için iki işlem vardır. Bu çalışmada kullanılan işlemler aşağıdaki gibidir:

- Birinci Metot;

1. Explorer simgesine tıkla
2. Filtreler simgesini seç
3. Unsupervised sekmesini seç
4. Parametreye sekmesini seç
5. Eksik Değerler ile Değiştir
6. Uygula
7. Düzenle

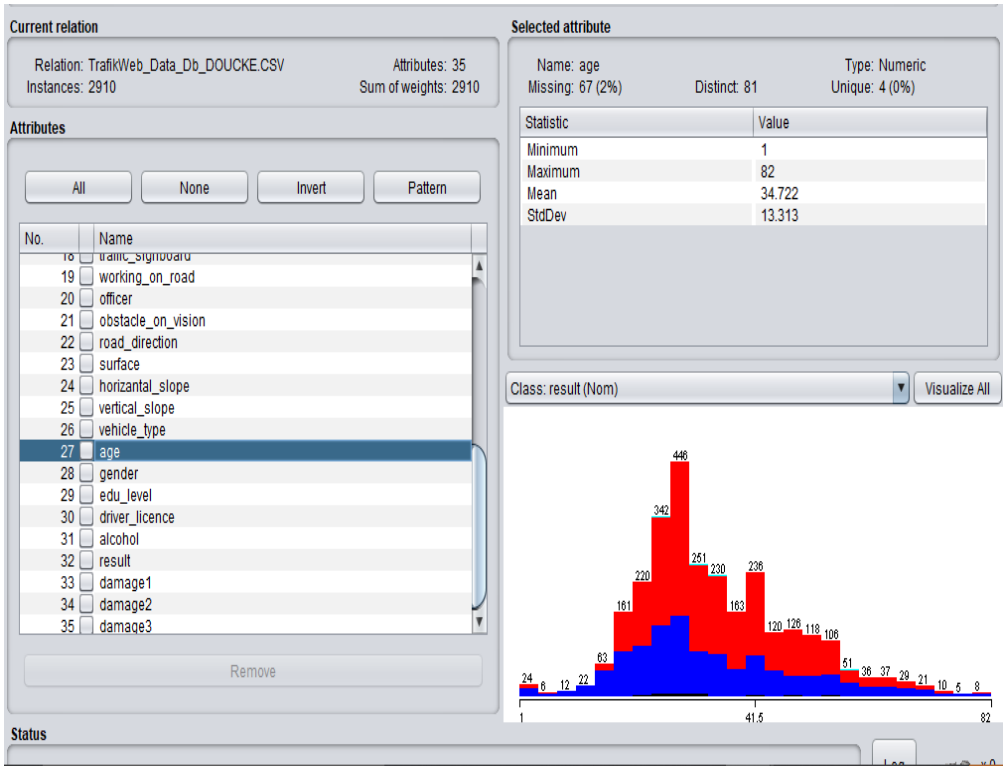
Eksik verileri görüntüle



Şekil 3.4. Eksik verilerin görüntülenmesi

- İkinci Metot

WEKA'da Ekle + Ekran Görüntüsü + Ekran kırkma düğmelerine tıklanarak, her bir özniteliğin hangi eksik değere sahip olduğu kullanıcı tarafından net bir şekilde görülebilir.



Şekil 3.5. Özniteliklerle ilgili eksik verilerin görselleştirilmesi ve doğrulanması

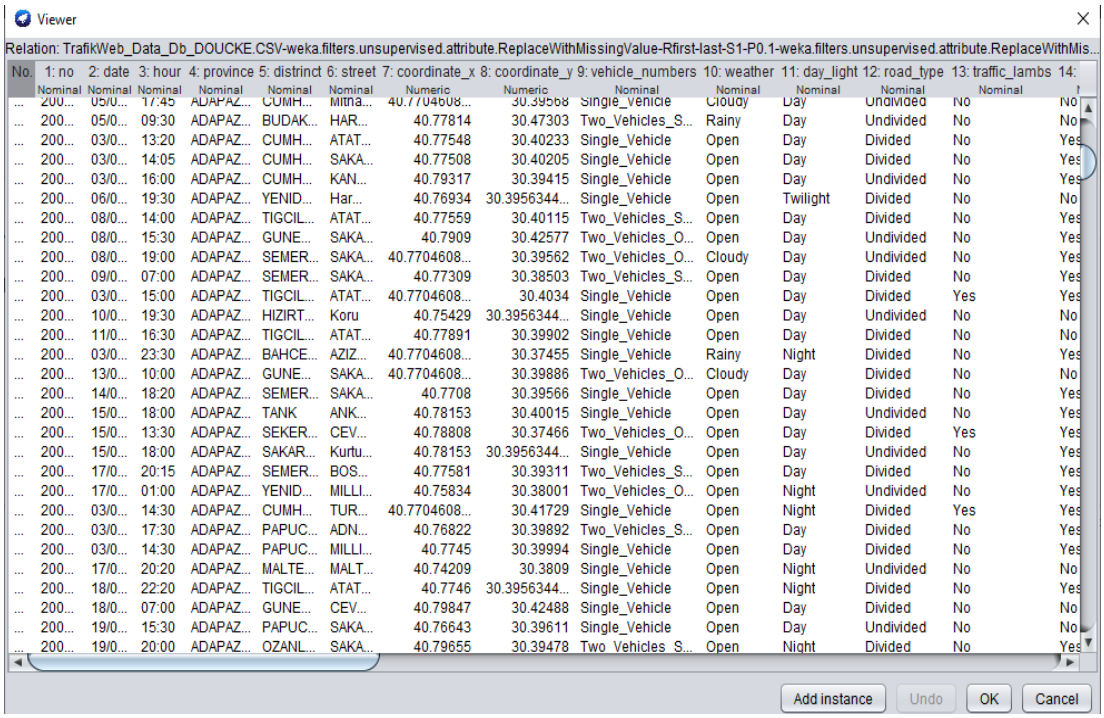
3.3.1.2.2. WEKA'da eksik değerlerin değiştirilmesi

Eksik değerleri değiştirmek için iki yöntem sunulmaktadır.

- Birinci Metot

1. Explorer simgesine tıkla
2. Filtreler simgesini seç
3. Unsupervised sekmesini seç
4. Parametreye sekmesini seç

5. Eksik Değerleri Değiştir
6. Uygula
7. Düzenle



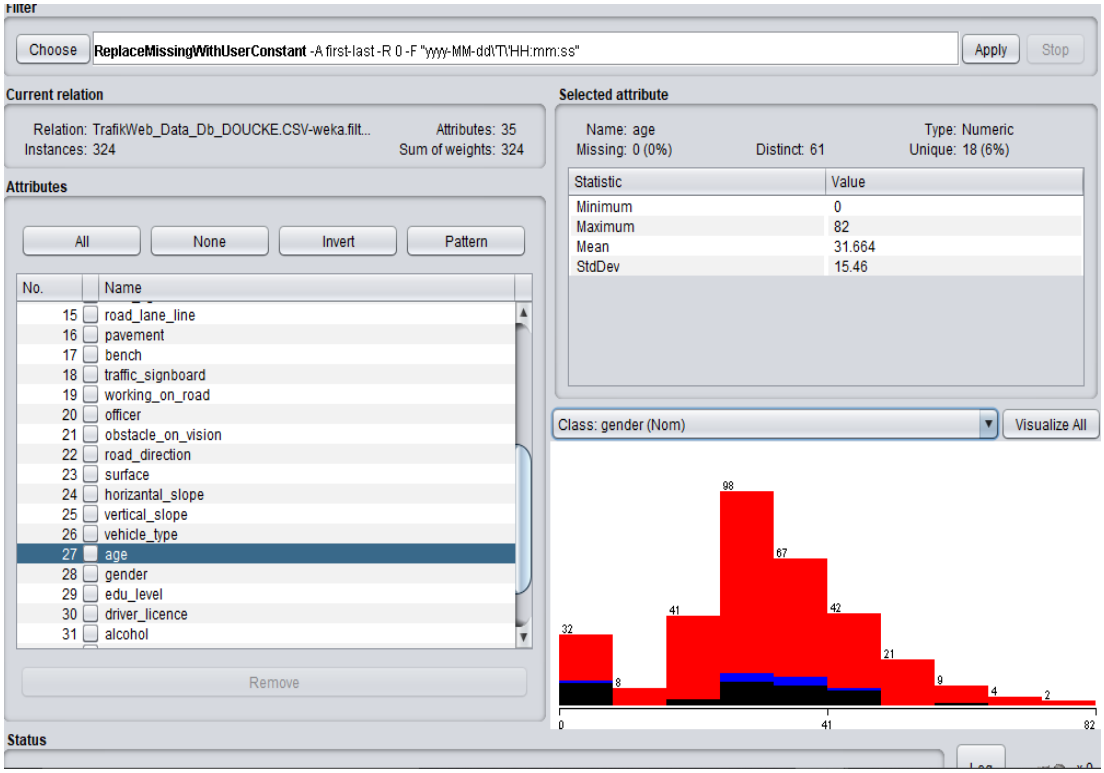
Şekil 3.6. Eksik verileri ekledikten sonra veri görselleştirme

Şekil 3.6.'da bu WEKA'da karar ağacının geliştirilmesinde çok önemli bir adım çünkü eksik veriye sahip özniteliklerimiz varsa, karar ağacı geliştirilemez, bu yüzden bu adım çok önemlidir

- İkinci Metot

1. Explorer simgesine tıkla
2. Filtreler simgesini seç
3. Unsupervised sekmesini seç
4. Parametreye sekmesini seç
5. Eksik Değerleri Değiştir
6. Uygula
7. Düzenle

Bazı özniteliklerin eksik değerleri yoktur, bu nedenle yalnızca eksik değerleri olan öznitelikleri değiştirilir.



Şekil 3.7. Eksik verilerini tamamlanması

3.3.1.2.3. Weka'da aykırı değerler ve aşırı değerlerin ayıklanması

- Aykırı Değerler

Aykırı değerler, diğer veri noktalarından uzak olan veri noktalarıdır. Başka bir deyişle, bir veri kümesindeki olağandışı değerlerdir. Aykırı değerler, testlerin önemli bulguları kaçırmasına veya gerçek sonuçları çarpıtmasına neden olabileceğinden, birçok istatistiksel analiz için sorunludur.

Ne yazık ki, aykırı değerleri kesin olarak belirlemek için açık ve kesin istatistiksel kurallar yoktur. Aykırı değerlerin bulunması, konu alanı bilgisine ve veri toplama sürecinin anlaşılmasına bağlıdır. Kesin bir matematiksel tanım olmasa da, aykırı adayları bulmak için kullanabileceğiniz kılavuzlar ve istatistiksel testler vardır.

Aykırı değerlerin kontrol edilmesinin nedenlerinden biri, analiz sürecine dahil edilecek verilerin kalitesini doğrulamaktır. Verilerdeki aykırı değerler için olası kaynaklardan biri doğru olmayan değerlerdir. "Yanlış değerler" için farklı iki potansiyel kaynak, eksik veri ve veri girişi veya kaydındaki hatalardır.

Değerlerin bilinmediği zamanlarda, verileri giren kişi bunu belirtmek için bir değer kullanabilir. Uygulamada karşılaşılan bazı durumlarla ilgili açıklamalar aşağıda verilmiştir.

Sayısal değerler: Beklenen değer aralığının dışında olduğu bilinen değerler varsa, bunlar eksik değerleri belirtmek için kullanılabilir. Örneğin sürücü yaşı verisi olarak şu değerler yanlışlıkla veri setine girilmiş olabilir.

- 0
- 9999
- -9999

Dize değerleri: Eksik veya bilinmeyen dize için genellikle yinelenen tek bir karakter, noktalama işareti veya belirli sözcükler kullanılabilir. Örnekler şunları içerir:

- xxxx, aaa, yy
- ., ,, ?, *
- Bilinmiyor, Belirtilmemiş, Eksik

Tarihler: Tarihler genellikle ya bir olayın tarihi ya da bir kişinin doğum tarihidir. Eksik değerler için gerçek tarih olamayacak tarihler kullanılabilir. İnsanlar için bu genellikle kişinin doğum tarihini imkânsız kılan tarihlerdir. Bunlar, örneğin 2019 yılında kazaya karışan bir sürücünün doğum tarihinin, kaza anından çok önceki veya çok ilerideki tarihleri olabilir. Bunlara örnek olarak aşağıdaki veri içerikleri verilebilir.

- 1850-01-01, 1900-01-01
- 2130-01-01, 3000-12-31

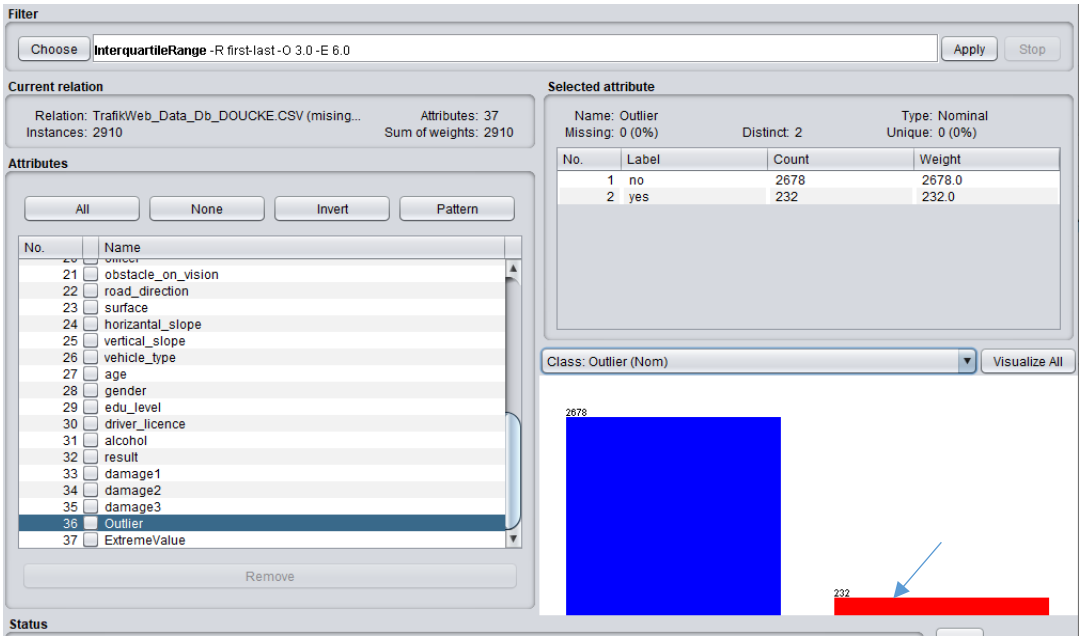
İsimler: Eksik isimler, yukarıda dizeler için belirtilenlere benzer yöntemlerle kodlanabilmekle beraber, sıklıkla kullanılan bazı ek kurallar vardır. Bazı kalıplar, uzun süredir adı bilinmeyen kişiler için kullanılmaktadır. İş alanına bağlı olarak genellikle başka genel terimler de kullanılabilir. Bazen eksik isimler, ad ve soyadların kombinasyonundan tespit edilebilmektedir. Eksik isim alanları, tek tek veya aşağıdaki kombinasyonları içerebilir:

- Cinsiyet,
- Eğitim düzeyi,
- Araç Plakası,
- Hava Durumu,
- Sokak,
- İlçe...

Sayısal değerler hariç diğer veri türlerinin tümü genellikle doğrudan sıralanamaz. Ancak, ilgili alanların gruplandırılmış bir sayımı tanımlanırsa, "varsayılan" değerleri belirlemek genellikle kolaydır.

Aykırı değerleri tespit etmek için şu adımlar takip edilir:

- Explorer simgesine tıkla
- Filtreler simgesini seç
- Unsupervised sekmesini seç
- Özelikle sekmesini seç
- InterquartileRange sekmisine tıkla
- Uygula

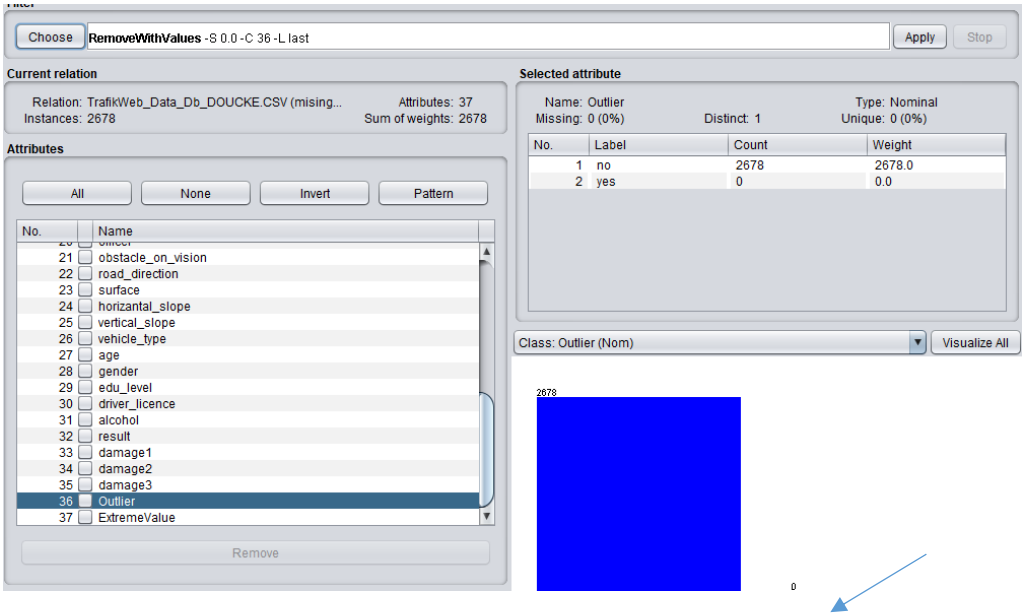


Şekil 3.8. Aykırı değerlerin görselleştirilmesi

Önişleme süreci gerçekleştirildikten sonra, tez konusuna ait verilerle ilgili olarak 232 aykırı değer tespit edilmiştir.

Belirlenen bu aykırı değerlerin ortadan kaldırılma işlemleri adım adım sırası ile,

- Choose simgesine seç
- Unsupervised sekmesine tıkla
- Instance sekmesine tıkla
- RemoteWithValues sekmesine tıkla
- Uygula



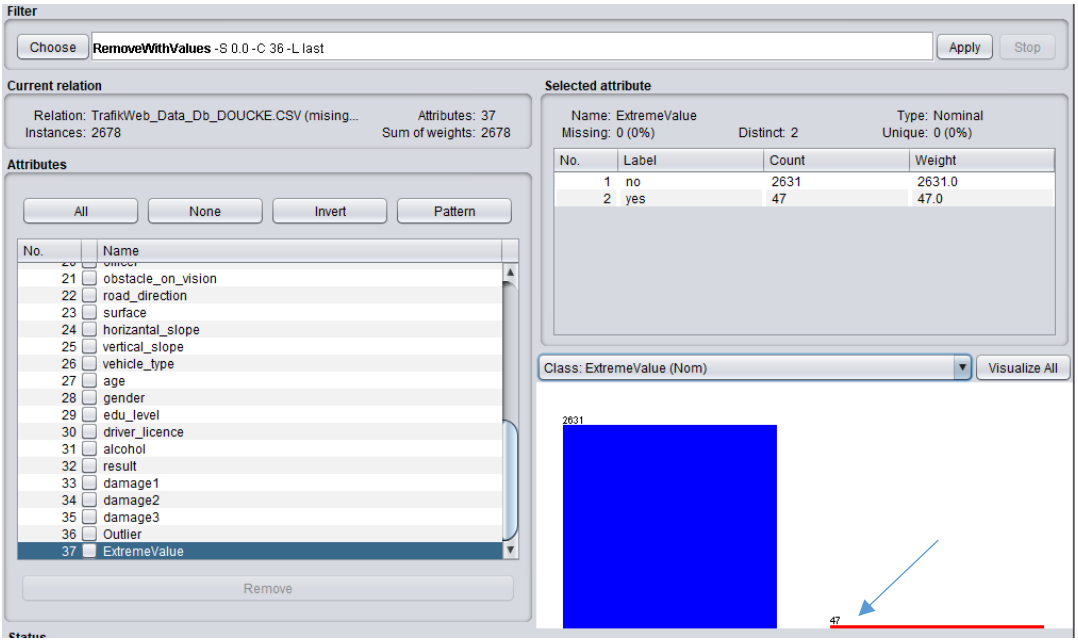
Şekil 3.9. Aykırı değerlerin silinmesinin görselleştirilmesi

- Aşırı değerler

Bu karakteristik değerler, en küçük (minimum) veya en büyük (maksimum) değerler olup uç değerler olarak bilinir. Örneğin, en küçük ve en uzun insanların vücut ölçüleri, insanların boy karakteristikleri için uç değerleri temsil eder.

Aşırı değerlerin tespiti şu adımlar üzerinden yapılır.

- Explorer simgesine tıkla
- Filtreler simgesini seç
- Unsupervised sekmesini seç
- Özelikle sekmesini seç
- InterquartileRange sekmisine tıkla
- Uygula

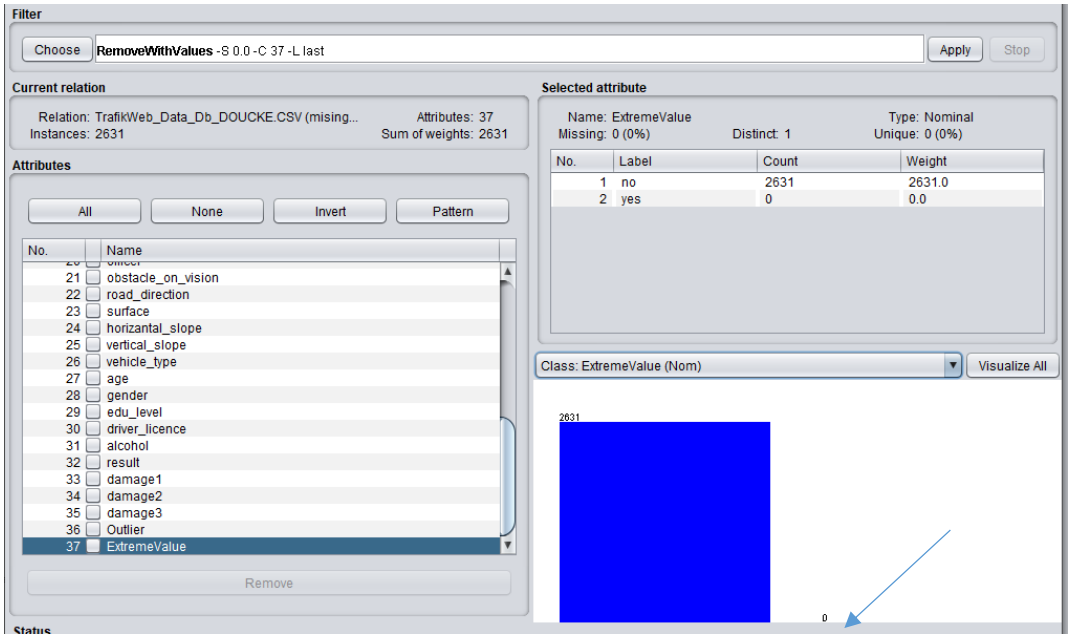


Şekil 3.10. Uç değerlerin görselleştirilmesi

İşlem teze konu veri seti için gerçekleştirildiğine, 47 aşırı değer belirlenmiştir.

Bu aşırı değerlerin ortadan kaldırılma adımları olarak aşağıdaki süreç takip edilir.

- Choose simgesine seç
- Unsupervised sekmesine tıkla
- Instance sekmesine tıkla
- RemoteWithValues sekmesine tıkla
- Uygula



Şekil 3.11. Aşırı değerlerin silinmesi

3.3.2. Sınıflandırma

Sınıflandırma, yapılandırılmış veya yapılandırılmamış veriler üzerinde gerçekleştirilebilir. Sınıflandırma, verileri belirli sayıda sınıfa ayırdığımız bir tekniktir. Bir sınıflandırma probleminin temel amacı, yeni bir verinin altına düşeceği kategoriyi/sınıfı belirlemektir.

Her bir hedef sınıfı belirlemek için kullanılabilecek daha iyi sınır koşulları elde etmek için eğitim veri seti kullanılır. Sınır koşulları belirlendikten sonraki görev, hedef sınıfı tahmin etmektir. Tüm süreç sınıflandırma olarak adlandırılır.

Makine öğreniminde sınıflandırma içeriğinde kullanılan terminolojik terimler aşağıda sunulmuştur.

- Sınıflandırıcı: Girdi verilerini belirli bir kategoriye eşleyen algoritma.
- Sınıflandırma modeli: Bir sınıflandırma modeli, yeni veriler için sınıf etiketlerini/kategorilerini tahmin ederek, eğitim için verilen girdi değerlerinden bazı sonuçlar çıkarmaya çalışır.

- Özellik: Bir özellik, gözlemlenen bir olgunun bireysel olarak ölçülebilir bir niteliğidir.
- İkili Sınıflandırma: İki olası sonucu olan sınıflandırma görevi. Örn: Cinsiyet sınıflandırması (Erkek / Kadın)
- Çok sınıflı sınıflandırma: İkiden fazla niteliği olan sınıflandırma. Çok sınıflı sınıflandırmada her numuneye yalnızca bir hedef etiket atanır. Örn: Bir hayvan kedi ya da köpek olabilir ama ikisi aynı anda olamaz.
- Çok etiketli sınıflandırma: Her örneğin bir dizi hedef etiketle (birden fazla sınıf) eşleştirildiği sınıflandırma görevi. Örneğin: Bir haber makalesi aynı anda spor, kişi ve konum hakkında olabilir.

Aşağıdakiler, bir sınıflandırma modeli oluşturmayla ilgili adımlardır:

- Kullanılacak sınıflandırıcıyı başlatın.
- Sınıflandırıcıyı eğitin: scikit-learn'deki tüm sınıflandırıcılar, verilen X eğitim verisi ve y eğitim etiketi için modele (eğitim) uyması için bir fit(X,y) yöntemini kullanır.
- Hedefi tahmin edin: Etiketlenmemiş bir X gözlemi verildiğinde, tahmin(X), tahmin edilen y etiketine döndürür.
- Sınıflandırıcı modelini değerlendirin

3.3.2.1. WEKA'da sınıflandırma algoritmaları türleri

WEKA, çok sayıda sınıflandırma algoritmasını kullanıma sunar.

WEKA platformunu kullanmanın faydalarından birisi, mevcut çok sayıda makine öğrenimi algoritmasını, makine öğrenimi problemlerinin çözümüne sunmasıdır

WEKA' nın kullandığı beş (5) sınıflandırma algoritmasının nasıl çalıştığı aşağıda açıklanarak, temel algoritma parametreleri vurgulanmış ve WEKA Explorer ara yüzünde gösterilmiştir.

İncelenecek 5 algoritma şunlardır:

- Lojistik regresyon
- Naive Bayes
- Karar ağacı
- K-En Yakın Komşular
- Vektör destek makineleri

Bunlar, başlangıç noktası olarak sınıflandırma problemi üzerinde denenebilecek 5 temel algoritmadır.

Girdi değişkenlerinin aynı ölçeğe sahip sayısal veriler olması, standart bir makine öğrenimi sınıflandırma problemi için geliştirilen algoritmaları etkin kılmaktadır. WEKA da bu sınıflandırma problemi işlem adımları şu adımlar üzerinden gerçekleştirilir.

- WEKA Explorer'ı başlatın:
- WEKA GUI Seçici'yi açın.
- WEKA Explorer'ı açmak için “Explorer” düğmesine tıklayın.
- İlgili veri kümesini data/veridosyası.arff dosyasından yükleyin.
- Sınıflandır sekmesini açmak için “Sınıflandır”a tıklayın.

1. Lojistik regresyon

Lojistik regresyon, ikili bir sınıflandırma algoritması olup, girdi değişkenlerinin sayısal ve Gauss (çan eğrisi) dağılımına sahip olduğunu varsayar. Verilerin Gauss dağılımına uygun olmaması durumda dahi, lojistik regresyon yine de iyi sonuçlar verebilmektedir.

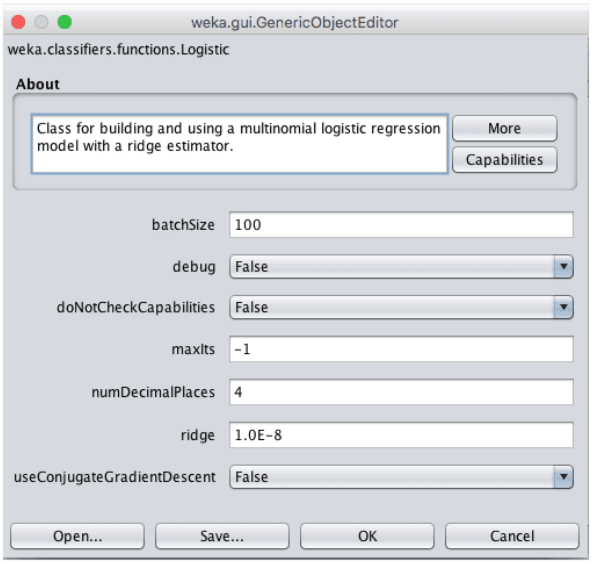
Algoritma, doğrusal olarak bir regresyon fonksiyonuna birleştirilen ve bir lojistik (s-şekilli) fonksiyon kullanılarak dönüştürülen her giriş değeri için bir katsayı öğrenir.

Lojistik regresyon, hızlı ve basit bir teknik olmakla beraber, bazı problemlerde çok etkili sonuçlar üretebilmektedir.

WEKA uygulaması çok sınıflı sınıflandırma problemlerini desteklemek için uyarlanmış olmasına rağmen, lojistik regresyon sadece ikili sınıflandırma problemlerini desteklemektedir. Bu algoritmanın uygulama adımları aşağıda verilmiştir.

Lojistik regresyon algoritmasını seçin:

- "Seç" düğmesine tıklayın ve "fonksiyonlar" grubu altında "Lojistik"i seçin.
- Algoritma yapılandırmasını gözden geçirmek için algoritmanın adına tıklayın.



Şekil 3.12. Regresyon Algoritması Penceresi

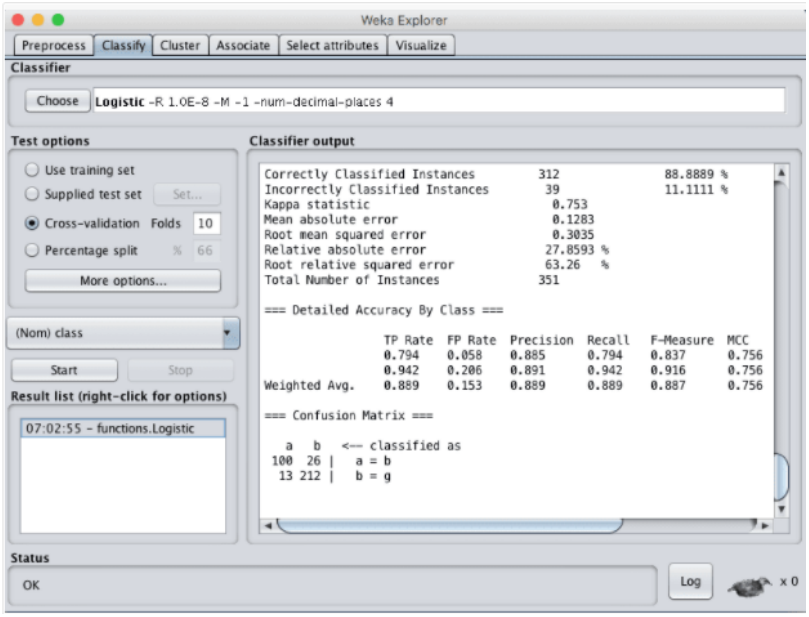
Algoritma, sabit sayıda yineleme (maxIts) için çalışabilir, ancak varsayılan olarak, algoritmanın yakınsaklığı elde etme aşamasına kadar çalışacaktır.

Uygulama, bir tür düzenlileştirme olan bir geri beslemeli tahmin mekanizması kullanır. Bu yöntem, model tarafından öğrenilen katsayıları en aza indirerek eğitim sırasında modeli basitleştirmeyi amaçlamaktadır. Bu prosedür, katsayıların boyutunu

azaltmak için algoritmaya ne kadar baskı uygulanacağını tanımlar. Bunu 0'a ayarlamak, bu düzenlemeyi kapatacaktır.

- Algoritma yapılandırmasını kapatmak için "Tamama tıklayın.
- Algoritmayı Ionosphere veri kümesinde çalıştırmak için "Başlat" düğmesine tıklayın.

Varsayılan konfigürasyonla, tez veri seti için lojistik regresyonun %88'lik bir doğruluk değeri üretmektedir.



Şekil 3.13. Regresyon algoritması penceresi

2. Naive Bayes

Naive Bayes bir sınıflandırma algoritmasıdır. Geleneksel olarak, sayısal girdilerin dağılımsa desteklenmesine rağmen, girdi değerlerinin nominal olduğunu varsayar.

Naive Bayes, her sınıf için önceki olasılığın eğitim verilerinden hesaplandığı ve birbirinden bağımsız olduğu varsayıldığı (teknik olarak koşullu bağımsız olarak bilinir) Bayes Teoreminin basit bir uygulamasını kullanır.

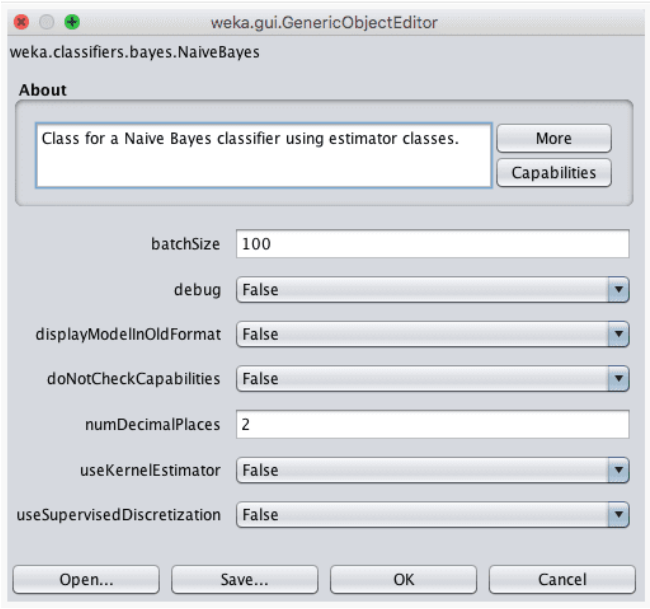
Bu gerçekçi olmayan bir varsayımdır, çünkü bu varsayım olasılıkların hızlı ve kolay hesaplanmasını sağlamasına rağmen, değişkenlerin etkileşime girmesi ve bağımlı olmaları gerçekliği göreceli olarak göz ardı edilebilmektedir. Bu gerçekçi olmayan varsayım altında bile, Naive Bayes 'in çok etkili bir sınıflandırma algoritması olduğu gösterilmiştir.[27]

Naive Bayes, her sınıf için sonsal olasılığı hesaplar ve en yüksek olasılığa sahip sınıf için bir tahmin yapar. Bu nedenle hem ikili sınıflandırma hem de çok sınıflı sınıflandırma problemlerini destekler.

Veri seti için Naive Bayes algoritmasını WEKA da çalıştırmak için şu adımlar uygulanır.

Naive Bayes algoritmasını seçin:

- "Seç" düğmesine tıklayın ve "bayes" grubu altında "Naive Bayes"i seçin.
- Algoritma yapılandırmasını gözden geçirmek için algoritmanın adına tıklayın.

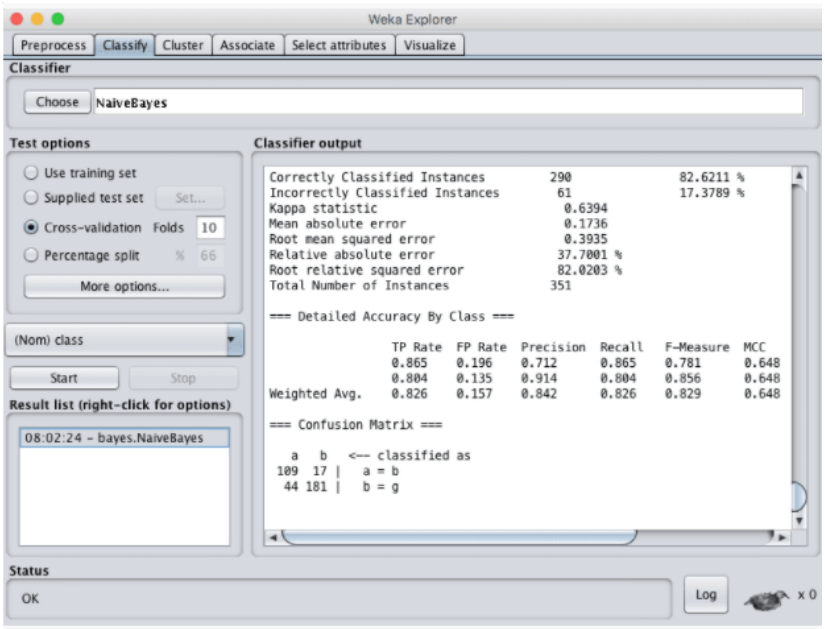


Şekil 3.14. Naive Bayes algoritmasını penceresi

Varsayılan olarak, her sayısal nitelik için bir Gauss dağılımı varsayılır. Algoritma, veri kümesindeki özniteliklerin gerçek dağılımıyla daha iyi eşleşebilecek useKernelEstimator argümanıyla bir çekirdek tahmincisi kullanacak şekilde değiştirilebilir. Alternatif olarak, useSupervisedDiscretization parametresiyle sayısal öznitelikler, nominal özniteliklere otomatik olarak dönüştürebilir.

- Algoritma yapılandırmasını kapatmak için "Tamam"a tıklayın.
- Algoritmayı veri kümesinde çalıştırmak için "Başlat" düğmesine tıklayın.

Aşağıdaki şekilden de görülebileceği gibi Naive Bayes'in algoritmasının veri setine uygulanması ile %82 doğruluk değerine ulaşılmıştır.



Şekil 3.15. Veri kümesine Naive Bayes algoritması uygulanarak elde edilen yüzde.

3. Karar ağaçları

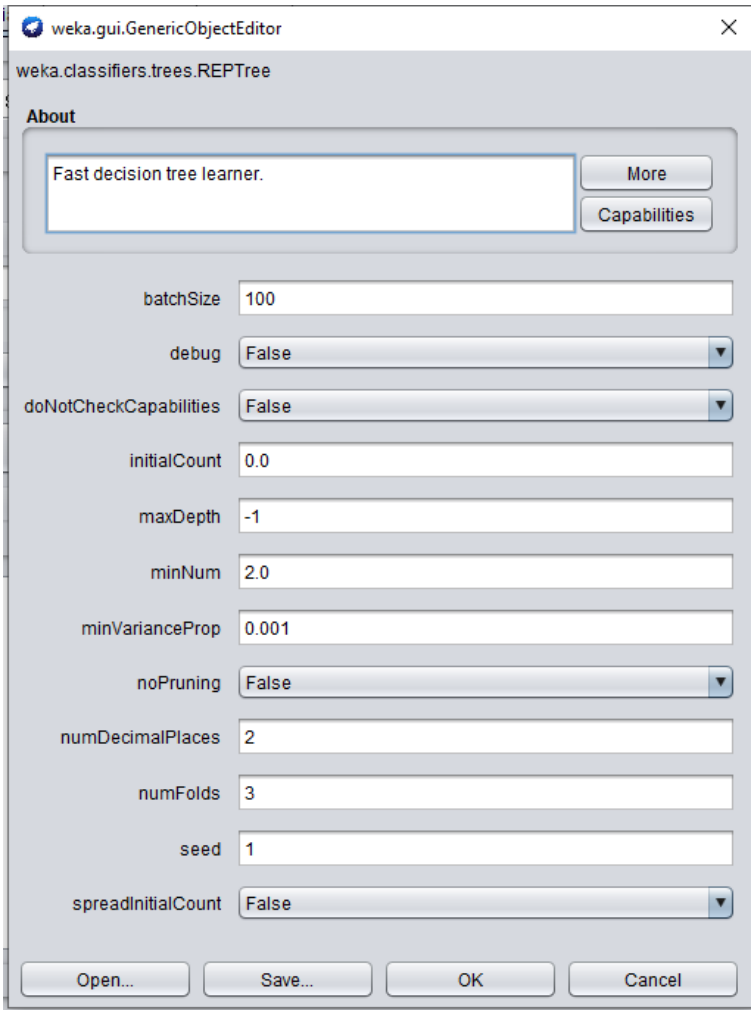
Karar ağaçları, sınıflandırma ve regresyon problemlerini desteklemektedir. Karar ağaçları yakın zamanda Sınıflandırma ve Regresyon Ağaçları (CART) olarak anılmaktaydı. Bir veri örneğini değerlendirmek için ağacın kökünden başlarlar ve bir tahmin yapılabilene kadar dallanarak bir ağaç yapısı oluşturarak, veriler arasındaki karşılıklı ilişki ve etkileşimi ortaya koyarlar. Bir karar ağacı oluşturma süreci,

tahminler yapmak için en iyi ayrım noktasının seçimi esasına dayanarak, sürecin tekrarlanması şeklinde çalışır.

Ağaç oluşturulduktan sonra, modelin yeni verilere genelleme yeteneğini geliştirmek için budama işlemi uygulanabilir.

Karar ağaçları algoritmasının işletim adımları aşağıda sunulmuştur. Karar ağacı algoritmasını seçin;

- "Seç" düğmesine tıklayın ve "ağaçlar" grubu altında "REPTree"yi seçin.
- Algoritma yapılandırmasını gözden geçirmek için algoritmanın adına tıklayın.



Şekil 3.16. Karar ağaçları penceresi

Ağacın derinliği otomatik olarak tanımlanabilse de maxDepth özniteliğinde bir derinlik de belirtilebilir.

NoPruning parametresini True olarak ayarlayarak budama kapatılabilir. Bununla beraber bu tercihte algoritmanın performansında düşüş söz konusu olabilir.

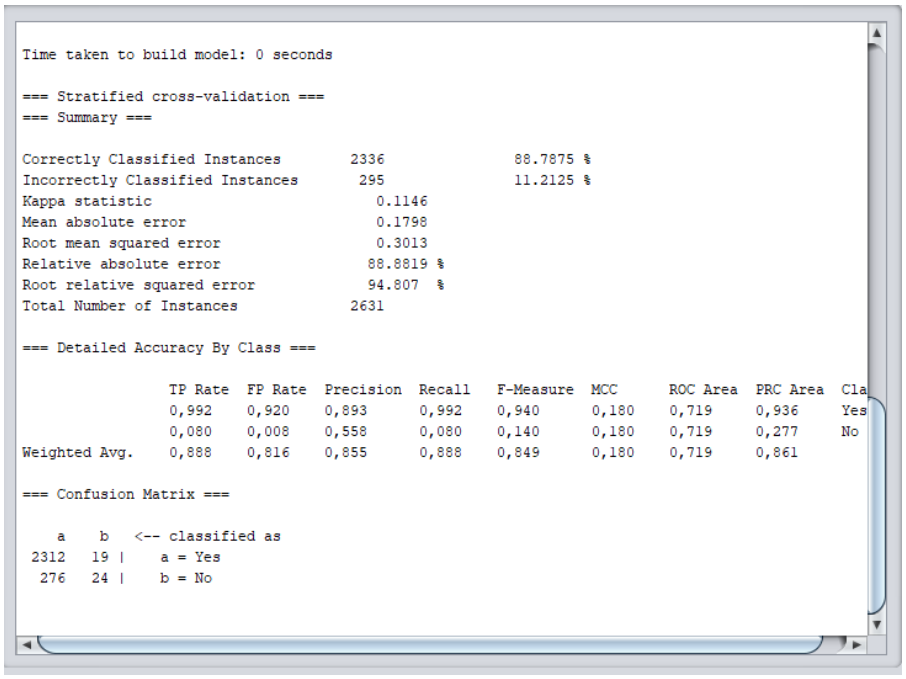
MinNum parametresi, eğitim verilerinden ağaç oluşturulurken bir yaprak düğümde ağaç tarafından desteklenen minimum örnek sayısını tanımlamaktadır.

Algoritmanın işlem adımları şu şekildedir.

Algoritma yapılandırmasını kapatmak için "Tamam"a tıklayın.

- Algoritmayı Ionosphere veri kümesinde çalıştırmak için "Başlat" düğmesine tıklayın.??????????????

Varsayılan yapılandırma ile karar ağacı algoritmasının %89 doğruluk derecesinde sonuçlar elde edilebilmiştir.

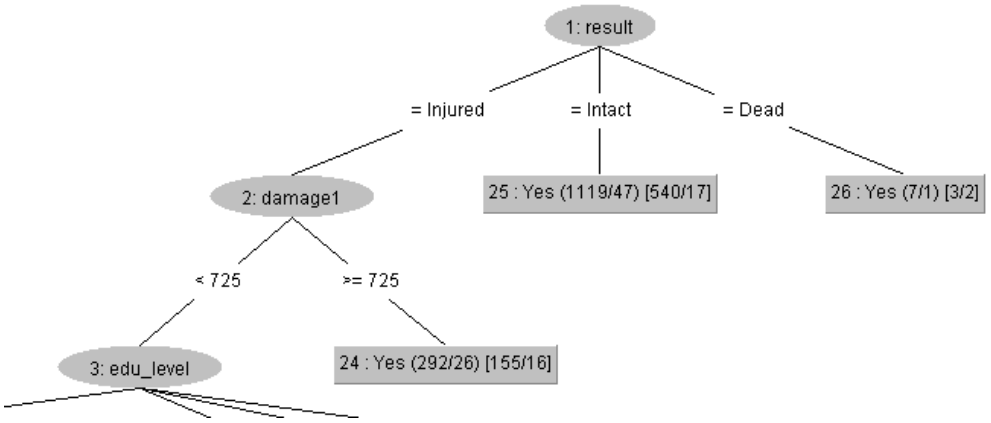


```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances        2336               88.7875 %
Incorrectly Classified Instances       295               11.2125 %
Kappa statistic                          0.1146
Mean absolute error                      0.1798
Root mean squared error                  0.3013
Relative absolute error                 88.8819 %
Root relative squared error             94.807  %
Total Number of Instances             2631

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Cla
                 0,992    0,920    0,893      0,992    0,940      0,180    0,719     0,936     Yes
                 0,080    0,008    0,558      0,080    0,140      0,180    0,719     0,277     No
Weighted Avg.    0,888    0,816    0,855      0,888    0,849      0,180    0,719     0,861

=== Confusion Matrix ===

    a    b   <-- classified as
 2312   19 |    a = Yes
  276   24 |    b = No
```

Şekil 3.17. Grafiği çizmeden önce verilerin karar ağacında görselleştirilmesi

WEKA da kullanılabilecek daha gelişmiş bir karar ağacı algoritması, J48 adı verilen C4.5 algoritmasıdır.

Karar ağaçları sonuçlarına ait örnek görsel aşağıda verilmiştir.



Şekil 3.18. Örnek karar ağacı grafiği

4. K-En Yakın Komşular

K-en yakın komşular algoritması hem sınıflandırmayı hem de regresyonu destekler. Kısaca kNN olarak da adlandırılır.

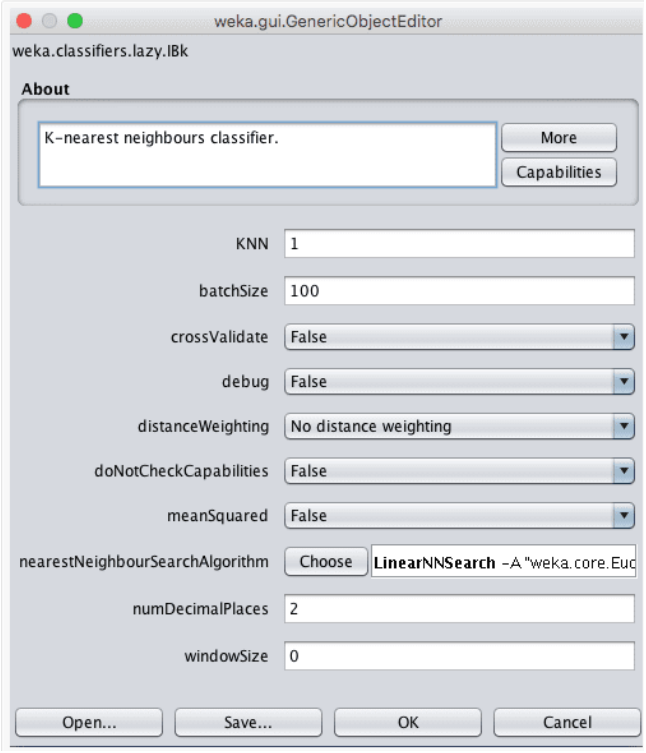
Tüm eğitim veri setini depolayarak ve bir tahmin yaparken de en benzer k eğitim modelini bulmak için sorgulayarak çalışır. Dolayısı ile ham eğitim veri seti üzerinden yapılan tek hesaplama, bir tahmin istendiğinde eğitim veri setinin sorgulanmasıdır.

Veri örnekleri arasındaki mesafelerle ilgili tahminlerde bulunması dışında problem hakkında çok fazla varsayımda bulunmayan bir algoritmadır. Bu nedenle, genellikle çok iyi performans gösterir.

KNN, sınıflandırma problemleriyle ilgili tahminlerde bulunurken, eğitim veri setindeki en benzer k örneğin modunu (en yaygın sınıf) kullanmaktadır.

WEKA da bu algoritmanın kullanımı şu adımlar üzerinden yapılmaktadır.

- “Seç” düğmesine tıklayın ve “Lazy” grubu altında “IBk” yi seçin.
- Algoritma yapılandırmasını gözden geçirmek için algoritmanın adına tıklayın.



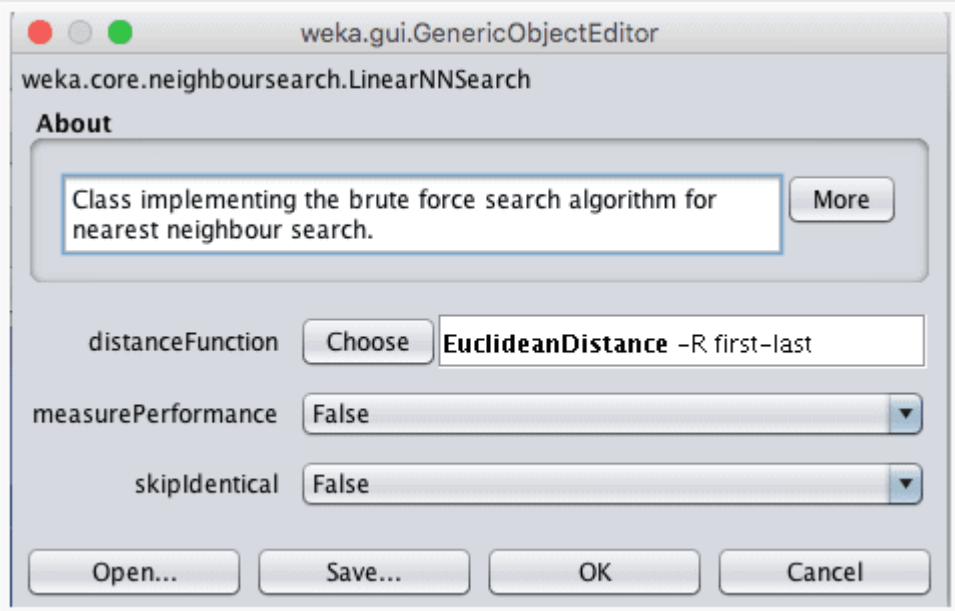
Şekil 3.19. K-en yakın komşular penceresi

Komşuluk kümesinin büyüklüğü k parametresi tarafından kontrol edilir.

Örneğin, k değeri 1'e ayarlanırsa, o zaman bir tahminin talep edildiği belirli bir yeni modele en çok benzeyen tek eğitim örneği kullanılarak tahminler yapılır. k için ortak değerler, daha büyük veri kümesi boyutları için daha büyük olan 3, 7, 11 ve 21'dir. WEKA, crossValidate parametresini “True” olarak ayarlayarak algoritma içindeki çapraz doğrulamayı kullanarak k için otomatik olarak uygun ve etkin bir değer belirleyebilmektedir.

Bir diğer önemli parametre ise kullanılan mesafe ölçüsüdür. Bu, eğitim verilerinin depolanma ve aranma şeklini kontrol eden en yakın NeighbourSearchAlgorithm'de yapılandırılmaktadır.

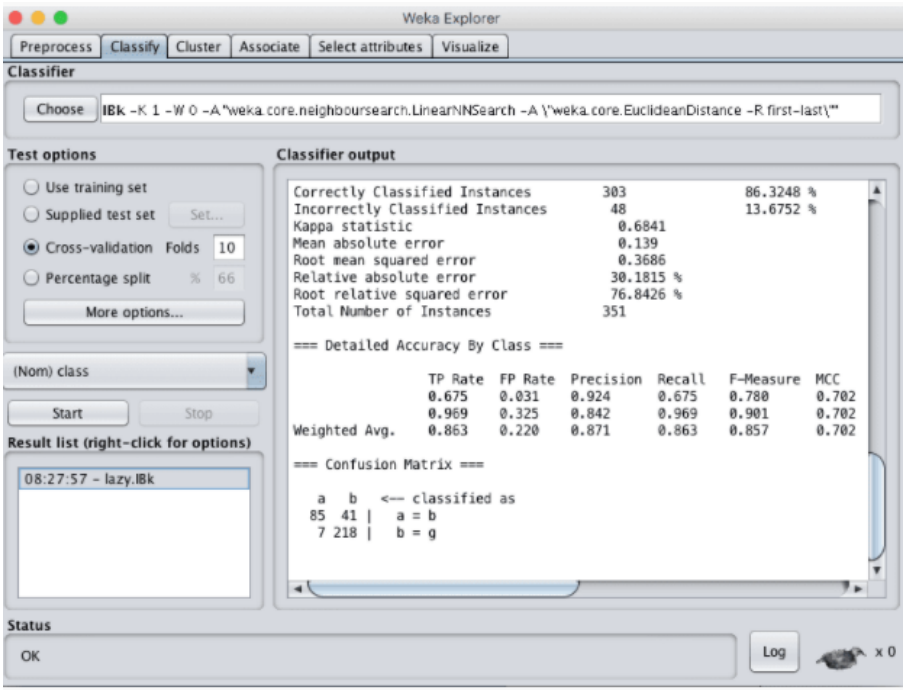
Mesafe ölçülendirilmesinde LinearNNSearch kullanılmaktadır. Bu arama algoritması aktive edildiğinde, bir DistanceFunction parametresinin seçilebildiği başka bir konfigürasyon penceresine erişim sağlanmaktadır. Varsayılan olarak, aynı ölçekte sayısal veriler için uygun örnekler arasındaki mesafeyi hesaplamak için Öklid mesafesi kullanılmaktadır. Veri setindeki özelliklerde ölçü veya tür olarak farklılık söz konusu olması durumunda Manhattan mesafesi kullanılabilir.



Şekil 3.20. Algoritma ile veri görselleştirme. K-En Yakın Komşular

Veri seti üzerinde bir dizi farklı k değerleri ve mesafe ölçüm kombinasyonları denemek ve hangisinin en iyi sonucu verdiğini tespit etmek algoritmadan beklenen etkin sonucun elde edilmesi için önerilen yöntemdir.

Algoritma yapılandırmasını kapatmak için "Tamam"a tıklayın. Algoritmayı Ionosphere veri kümesinde çalıştırmak için "Başlat" düğmesine tıklayın. Veri seti ve mevcut konfigürasyon ile kNN algoritması %86 doğruluk değerinde sonuçlar üretmektedir.



Şekil 3.21. kNN algoritması ile veri görselleştirme

5. Vektör destek makineleri

Vektör Destek Makinesi, ikili sınıflandırma problemleri için geliştirilmiş olmakla beraber, çok sınıflı sınıflandırma ve regresyon problemlerini desteklemek için de tekniğe uzantılar yapılmıştır. Algoritma genellikle kısaca SVM olarak adlandırılır.

SVM, işlem adımlarında sayısal girdi değişkenleri kullandığı için nominal değerleri otomatik olarak sayısal değerlere dönüştürür. Giriş verileri de kullanılmadan önce normalleştirilir.

SVM, verileri iki gruba en iyi şekilde ayıran bir hat bularak çalışır. Bunu da, eğitim veri kümesindeki sınıfları en iyi ayıran çizgiye en yakın olan veri örneklerini dikkate alan bir optimizasyon işlemi kullanarak yapar. Tekniğin adı örneklere destek vektörleri denilmesi ile ilişkilidir.

Hemen hemen tüm veri ilişkilendirme problemlerinde, sınıfları düzgün bir şekilde ayırmak için bir hat-çizgi çizilemez. Bu nedenle kısıtlamayı gevşetmek için çizginin

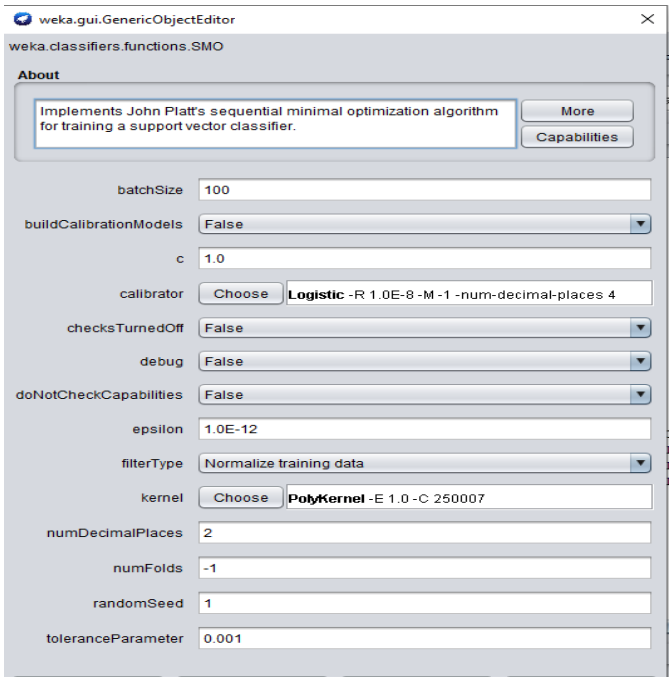
etrafına bir kenar boşluğu eklenir. Bazı örneklerin yanlış sınıflandırılmasına neden olmakla beraber, genel olarak iyi sonuçlar elde edilebilmektedir.

Birkaç veri kümesi yalnızca düz bir çizgiyle ayrılabilir. Bazen eğrileri ve hatta çokgen bölgeleri olan bir çizginin işaretlenmesi gerekir. Bu çizgileri çizmek ve tahminlerde bulunmak için verileri daha yüksek boyutlu bir uzaya yansıtarak SVM ile elde edilir. Projeksiyonu ve sınıfları ayırmadaki esneklik miktarını kontrol etmek için farklı çekirdekler kullanılabilir.

WEKA`da vektör destek makineleri algoritmasını aktive etmek için;

SVM algoritmasını seçin:

- "Seç" düğmesine tıklayın ve "fonksiyon" grubu altında "SVM"yi seçin.
- Algoritma yapılandırmasını gözden geçirmek için algoritmanın adına tıklayın.



Şekil 3.22. Vektör destek makineleri penceresi

WEKA'da karmaşıklık parametresi olarak adlandırılan C parametresi, sınıfları ayırmak için çizgi çizme sürecinin ne kadar esnek olabileceğini kontrol etmektedir. 0

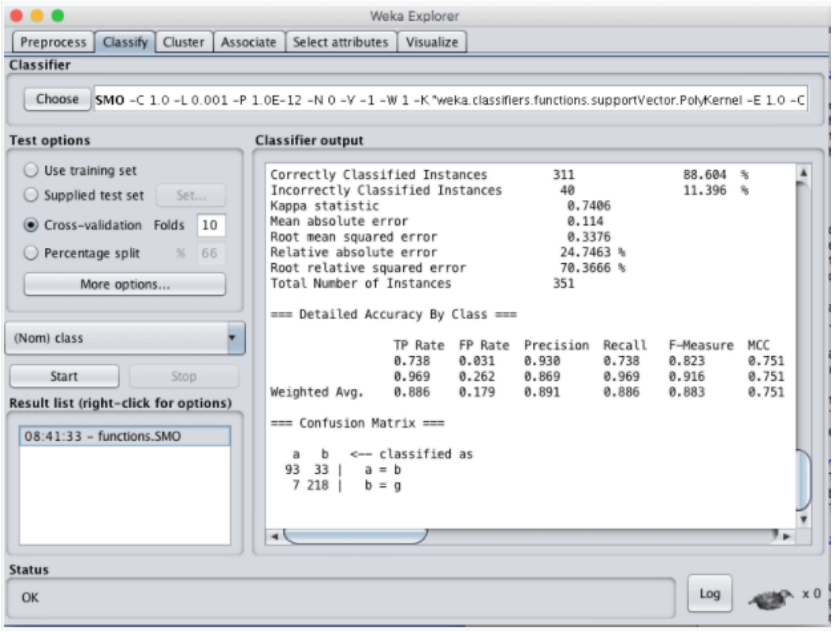
değeri, marjın ihlal edilmesine izin vermezken, algoritmada varsayılan değer 1 olarak kullanılmaktadır.

SVM'deki önemli bir parametre, kullanılacak Çekirdek türüdür. En basit çekirdek, verileri düz bir çizgi veya hiper düzlem ile ayıran bir doğrusal çekirdektir. Weka'daki varsayılan, sınıfları eğri veya hareketli bir çizgi kullanarak ayıracak bir Polinom Çekirdeğidir. Polinom derecesi ne kadar yüksek olursa, verilerin sınıflandırılması o kadar gerçekçi ve doğru sonuçlar üretecektir.

Sınıflandırmada kullanılan popüler ve güçlü çekirdekler, kapalı çokgenleri ve karmaşık şekilleri öğrenebilen RBF Çekirdeği veya Radyal Temel İşlev Çekirdeğidir.

Veri setleri üzerinde analiz yaparken, farklı çekirdekler ve C (karmaşıklık) değerlerinden oluşan farklı eşleşmeleri denemek ve hangisinin en iyi sonucu verdiğini görmek önerilen bir uygulamadır.

Aşağıdaki şekilden de görülebileceği üzere, SVM algoritması veri setimiz analizinde %88 doğruluk değerli sonuçlar üretmiştir.



Şekil 3.23. Weka'da Vektör destek makinesi algoritması ile veri görselleştirme

Yukarıda tez veri setine ait problem analizinde kullanılabilecek en iyi 5 sınıflandırma algoritmasının yapıları hakkında genel bilgiler verilmiştir.

3.4. Makine Öğrenmesinde Performans Ölçütleri

Makine öğrenimi modelleri oluşturma kavramı, yapıcı geri bildirim ilkesiyle çalışır. Bir model oluşturuyoruz, metrikler hakkında geri bildirim alıyoruz, iyileştirmeler yapıyoruz ve istenen kesinliği elde edene kadar devam ediyoruz. Derecelendirme metrikleri, bir modelin performansını açıklar. Değerlendirme metriklerinin önemli bir yönü, model sonuçları arasında ayrım yapabilme yetenekleridir.

BÖLÜM 4. ARAŞTIRMA BULGULARI

Çalışmamızda karar ağacı J48 algoritması ile temsil edilecektir. Aşağıdaki Tablo 4.1.'de, veri setinin özetini göstermektedir.

Tablo 4.1. Veri seti öz nitelik tablosu

Değişken	Açıklama	Değişken Tipi	Kategori
Y (Bağımlı değişken)	Sonuç	Üçlü	1. Ölü 2. Yaralı 3. Maddi hasarlı
X(Bağımsız değişken)	Kazaya Karışan Araç Sayısı	Dörtlü	0. Karşıdan gelen iki araç 1. Tek araç 2.Aynı yöne giden iki araç 3. Çoklu araç
	Hava Durumu	Üçlü	0. Açık 1. Bulutlu 2. Yağmurlu
	Gün Işığı	Üçlü	0.Gece 1. Gün 2. Alacakaranlık
	Yol Tipi	İkili	0. Bölünmüş 1. Bölünmüş
	Trafik Işıkları	İkili	0. Hayır 1. Evet
	Yol Işıklandırması	İkili	0. Hayır 1. Evet
	Yol şerit çizgisi	İkili	0. Hayır 1. Evet
	Kaldırım	İkili	0. Hayır 1. Evet
	Bank	İkili	0. Hayır 1. Evet
	Trafik İşaret Panosu	İkili	0. Hayır 1. Evet
	Yol yönü	İkili	0. Tek yön 1. İki yönlü
	Yüzey	İkili	0. Kuru 1. Islak

Tablo 4.1. (Devamı)

	Araç tipi	Sekizli	0. Otomobil 1. Kamyon 2. Motosiklet 3. Minibüs 4. Bisiklet 5. Otobüs 6. Diğer 7. Çekici
	Yaş	Üçlü	0. Genç 1. Orta Yaşlı 2. Yaşlı
	Cinsiyet	İkili	0. Kadın 1. Erkek
	Eğitim Seviyesi	Sıralı	1. Lise 2. Üniversite 3. İlk-Orta okul
	Sürücü ehliyeti	İkili	0. Hayır 1. Evet
	Alkol	İkili	0. Hayır 1. Evet

4.1. Karar Ağacı

Karar Ağacı, kök düğüm, dal (kenar veya bağlantı) ve yaprak düğüm olmak üzere üç bileşenden oluşan sınıflandırma tekniğidir. Kök, farklı nitelikler için test koşulunu temsil ederken, dal; testte olabilecek tüm olası sonuçları temsil etmekte, yaprak düğümler, ait olduğu sınıfın etiketini içermektedir. Kök düğüm, ağacın tepesi olarak da adlandırılan ağacın başlangıcındadır.

Sınıflandırma için J48, Reptree, Random Tree ve Random Forest algoritmaları birçok veri madenciliği analizlerinde yaygın olarak kullanılmaktadır. Bunlar, tümevarım yoluyla öğrenmenin bir biçimi olarak böl ve yönet stratejilerini kullanan karar ağaçlarıdır. Bu nedenle, bu algoritmalar, birbirine bağlı bir dizi düğümde hiyerarşik olarak yapılandırılan veri kümelerinde örüntü sınıflandırmasına yardımcı olan bir ağaç temsilini kullanır. Dahili düğümler, bir karar sabitiyle ilgili olarak bir girdi özniteliğini/özelliğini test eder ve bu şekilde bir sonraki azalan düğümün ne olacağını belirler. Bu nedenle, yaprak olarak kabul edilen düğümler, kendilerine ulaşan örnekleri ilişkili etikete göre sınıflandırır.

Karar ağacımızı oluştururken üç ana formül kullandığımızı bildirmek isteriz:

Entropi: Bir dizi verinin saflığını veya rastgeleliğini ölçmek için kullanılır.

$$Entropy(x) = \sum \ (P(x = k)) * log_2(P(x = k)) \quad (4.1)$$

Bilgi Kazancı: Bilgi kazancı açısından kök düğüm görevi gören en iyi özelliği bulmak için önce her bir tanımlayıcı özelliği kullanır ve veri kümesini bu tanımlayıcı özelliklerin değerlerine göre böleriz, ardından veri kümesinin l entropisini hesaplarız. Bu bize veri setini özellik değerlerine göre böldükten sonra kalan entropiyi verir. Ardından, bu varlık bölümünün bir varlığın bilgi kazancını veren orijinal entropiyi ne kadar azalttığını görmek için bu değeri veri kümesinin kökeninde hesaplanan entropiden çıkarırız ve aşağıdaki gibi hesaplanır:

$$InformatıonGain(feature) = Entropy(Dataset) - Entropy(feature) \quad (4.2)$$

Gini indeksi: Her sınıfın olasılıklarının karelerinin toplamı birden çıkarılarak hesaplanır. Bilgi kazanımı, farklı değerlere sahip daha küçük bölümleri tercih ederken, uygulanması kolay daha büyük bölümleri tercih eder.

$$Gini\ Index = 1 - \sum\big(P(x = k)\big)^2 \quad (4.3)$$

Bu tez çalışmasında karar ağacı analiz sürecinde detayları aşağıda sunulan J48 kullanılmıştır.

4.1.1. RepTree

RepTree algoritması hızlı bir karar ağacı öğrenicisi olup C4.5 sınıflandırma yazılımını içinde barındırmaktadır. C4.5 algoritması belli örneklem verisi üzerinden karar üretilmesi için karar ağacı sınıflandırması fonksiyonunu görmektedir. RepTree regresyon ağaç mantığını kullanarak, farklı iterasyonlarda çoklu ağaçlar üretebilmektedir. Üretilen bu üretilen ağaçlar içerisinden en iyi olanını seçmektedir. Bu da veri setinin temsilcisi olarak kabul edilmektedir. Ağacın çok fazla genişlemesini

önlemede kullanılan budama sisteminde, ağaç tarafından tahmin edilen ortalama karesel hata değeri ölçü olarak kullanılmaktadır.

Azaltılmış Hata Budama Ağacı ("REPT") temelde hızlı bir karar ağacı öğrenmesi olup elde edilen bilgiye dayanarak veya varyansı azaltarak karar ağacı oluşturmaktadır. [21]

4.1.2. J48 algoritması

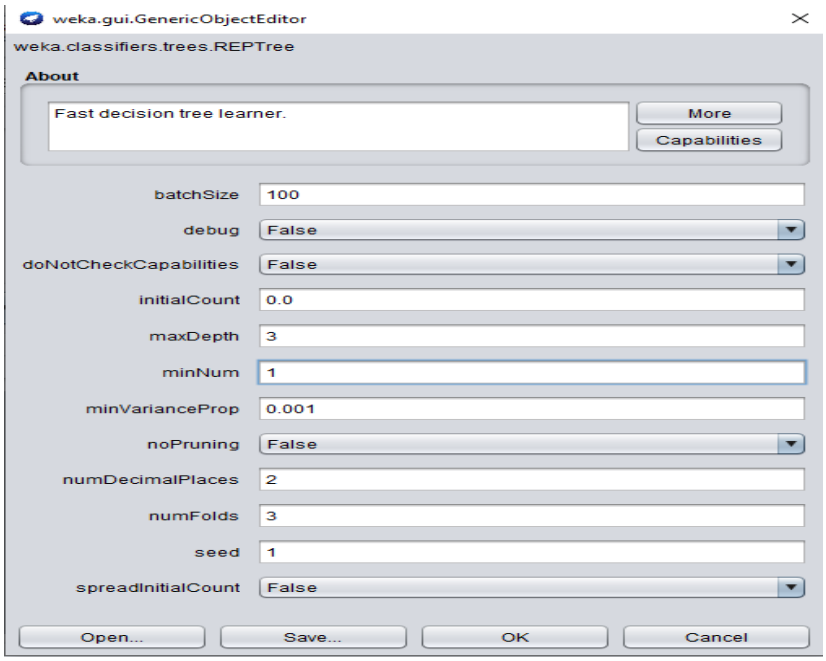
J48, WEKA veri madenciliği aracındaki C4.5 algoritmasının açık kaynaklı bir Java uygulamasıdır. İkili bir ağaç oluşturan bu yöntem, sınıflandırma problemleri için en kullanışlı karar ağacı yaklaşımlarından biridir. Sınıflandırma sürecini modellemek üzere bir karar ağacı oluşturarak tüm dallarda yukarıdan aşağıya doğru olası öznitelik etkileşimlerini tespit etmeye çalışır. Yeni bir öğeyi sınıflandırmak için öncelikle mevcut eğitim verilerinin öznitelik değerlerine dayalı bir karar ağacı oluşturulması gerekir. Bu nedenle, bir dizi öğeyle (eğitim seti) karşılaştığında, çeşitli örnekleri daha açık bir şekilde ayırt eden özelliği tanımlar. Veri örneklerinin en iyi şekilde sınıflandırabilmesi için daha fazla bilgi verebilen bu özelliğin, en yüksek bilgi kazancına sahip olduğu ifade edilmektedir. Bu özelliğin olası değerleri arasında, belirsizliği olmayan, yani kendi kategorisine giren veri örneklerinin hedef değişken için aynı değere sahip olduğu herhangi bir değer varsa, o dal sonlandırılır ve değer atanır. Diğer durumlar için en yüksek bilgi kazancını sağlayan başka bir öznitelik seçilir. Hangi nitelik kombinasyonunun belirli bir hedef değeri verdiği konusunda net bir karar alınana veya tüm nitelikler tamamlanana kadar süreç bu şekilde devam eder. Tüm özniteliklerin bitmesi veya mevcut bilgilerden kesin sonuç alınamaması durumunda, bu dalın altındaki öğelerin çoğunluğunun sahip olduğu bir hedef değeri bu şubeye atanır. Artık karar ağacımız olduğuna göre, ağaç için elde ettiğimiz öznitelik seçim sırasını takip edebiliriz. Karar ağacı modelinde görülenlerle ilgili tüm öznitelikler ve ilgili değerler kontrol edilerek, yeni örneğin hedef değeri tahmin edilebilir. J48 sınıflandırması, karar ağaçlarına veya onlardan oluşturulan kurallara dayanmaktadır.

4.2. Karar Ağacı Parametrelerinin İyileştirilmesi

Bu adımda, karar ağacı grafiğini geliştirmek için hangi algoritmanın seçilmesi gerektiğini belirlemek adına RepTree algoritması ve J48 algoritması ile daha yüksek bir Correctly Classified Instances yüzdesinin aranması işlemi yapılmaktadır.

4.2.1. RepTree algoritması parametreleri

Aşağıdaki şekilde belirtilen parametrelerden maxDepth ve minNum parametrelerine ait değerler kombinasyonlu bir şekilde değiştirilerek, Correctly Classified Instances değerleri artırılmaya çalışılır.



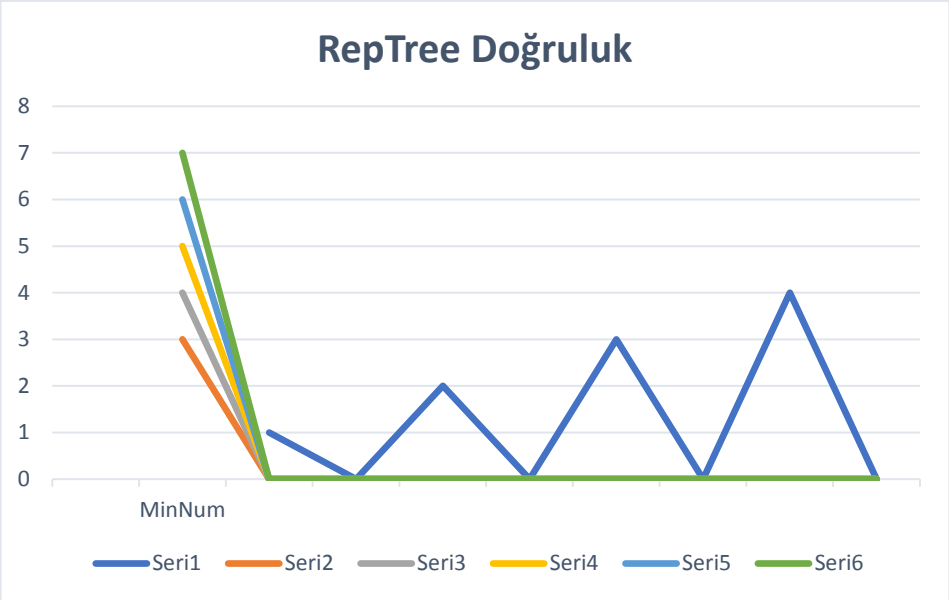
Şekil 4.1. Reptree Algoritması Parametreleri

4.2.1.1. RepTree doğruluk değerleri

Aşağıdaki Tablo 4.2. RepTree algoritması ile elde edilen Correctly Classified Instances yüzdelerine ait değerleri ifade etmektedir. Correctly Classified Instance yüzde değerinin MinNum`ın 4 ve maxDepht`in de 3 olduğu durumda en yüksek değer olan 75.0665 olarak elde edildiği görülmektedir.

Tablo 4.2. RepTree Correctly Classified Instances değerleri

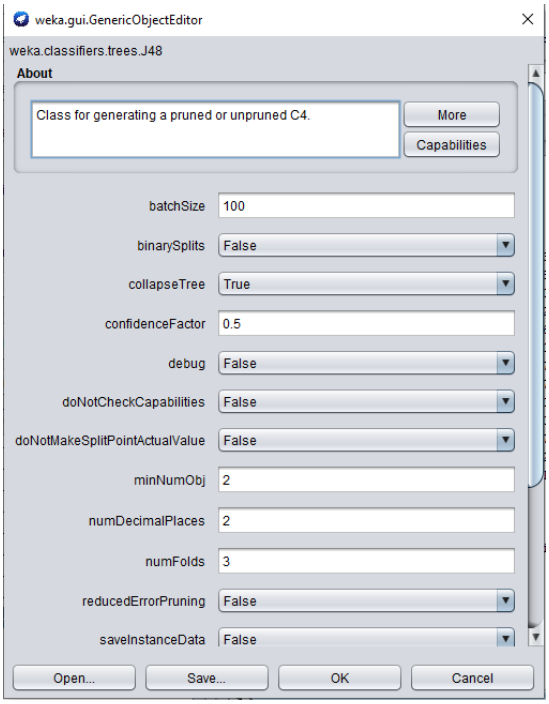
	MaxDepht					
MinNum		3	4	5	6	7
	1	74.8765	74.7244	74.6484	74.5344	74.3824
	1.5	74.8765	74.7244	74.6484	74.5344	74.3824
	2	74.8765	74.6104	74.5344	74.4204	74.2303
	2.5	74.8765	74.6104	74.5344	74.4204	74.2303
	3	74.8385	74.5724	74.4964	74.4584	74.3824
	3.5	74.8385	74.5724	74.5724	74.4584	74.3824
	4	75.0665	74.7624	74.7244	74.6864	74.6104
	4.5	75.0665	74.7624	74.7244	74.6864	74.6104



Şekil 4.2. Reptree doğruluk skoru

4.2.2. J48 Algoritması Parametreleri

Aşağıdaki şekilde belirtilen parametrelerden confidenceFactor ve minNumObj ve numFolds parametrelerine ait değerler kombinasyonlu bir şekilde değiştirilerek, Correctly Classified Instances değerleri artırılmaya çalışılır.



Şekil 4.3. J48 Algoritması Parametreleri

4.2.2.1. J48 doğruluk değerleri

Aşağıdaki Tablo 4.3. J48 algoritması ile elde edilen değerleri göstermektedir. Tablodan da görülebileceği üzere en yüksek Correctly classified Instances yüzdesel değeri ConfidenceFactor 0.1, maxDepht 2 ve Num Folds 3 iken elde edilen 75.6366 değeridir.

Tablo 4.3. J48 Correctly Classified Instances değerleri

ConfidenceFactor	MinNum obj	NumFolds	Correctly Classifier Instances
0.1	2	3	75.6366
0.2	2	4	75.4086
0.3	2	5	75.1045
0.4	2	6	74.5724
0.5	2	7	73.8502
0.7	2	8	71.2657
0.8	2	9	71.2657
0.9	2	10	71.2657
0.10	2	11	75.6366
0.11	2	12	75.5606
0.12	2	13	75.3706
0.13	2	14	75.3706
0.14	2	15	75.3706
0.15	2	16	75.3706

4.2.3. Reptree ile J48 sonuçlarının karşılaştırılması

RepTree algoritması ve J48 algoritması ile elde edilen en yüksek Correctly Classified Instances yüzdesel değerleri karşılaştırıldığında, ki bu değerler sırası ile 75.0665 ve 75.6366 olarak elde edilmişti, J48 algoritmasının daha iyi sonuç üretmesi dolayısı ile karar ağaçları oluşturulurken bu algoritma tercih edilmiştir.

Tablo 4.4. Reptree ve J48 Correctly Classified Instances değerlerinin karşılaştırılması

RepTree Doğruluk Skoru						
MaxDepht						
MinNum		3	4	5	6	7
	4	75.0665	74.7624	74.7244	74.6864	74.6104
J48 Doğruluk Skoru						
ConfidenceFactor	MinNum obj		NumFolds		Correctly Classifier Instances	
0.1	2		3		75.6366	

4.2.4. Sınıflara göre detaylı duyarlılık

- TP oranı (TP Rate): Gerçek pozitiflerin oranı (belirli bir sınıfta doğru şekilde sınıflandırılan örnekler).
- FP oranı (FP Rate): yanlış pozitiflerin oranı (belirli bir sınıfta yanlış sınıflandırılan örnekler)
- Kesinlik (Precision): Bir sınıfa ait doğrulanmış örneklerin o sınıfa ait toplam sınıflandırılmış örneklere oranı.
- Hatırlatma (Recall): Belirli bir sınıfta sınıflandırılan örneklerin bu sınıftaki gerçek toplama oranı (TP oranına eşdeğer)
- Ölçü F (F-Measure): Kesinlik ve hatırlatma için birleştirilmiş ölçüm değeri
- MCC: Makine öğreniminde ikili (iki sınıflı) sınıflandırmaların kalitesinin bir ölçüsü olarak kullanılır. Doğru ve yanlış pozitifleri ve negatifleri hesaba katar ve genellikle sınıflar çok farklı boyutlarda olsa bile kullanılabilecek dengeli bir ölçü olarak kabul edilir.
- ROC (Alıcı Çalışma Karakteristiği) Alan Ölçümü: WEKA'nın ürettiği en önemli değerlerden biridir. Sınıflandırıcıların genel olarak nasıl performans gösterdiği hakkında fikir verirler.

- ROC Alanı: ROC eğrisinin altında kalan alandır. WEKA, ROC eğrisinin altındaki bu alanı hesaplayıp yazdırmaktadır
- PRC Alanı: Bu bir "hassas geri çağırma eğrisi sınıflandırma ağacıdır

Aşağıdaki Tablo 4.5., karar ağacı grafiğinin çizilmesinden önce J.48 algoritmasının kullanılmasının uygun bir yaklaşım olacağına dair gerekli ve geçerli verileri sunmaktadır.

Tablo 4.5. J48 Algoritması ayrıntılı doğruluk parametreleri

TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Sınıf
0,489	0,085	0,768	0,489	0,597	0,460	0,731	0,651	Yaralı
0,916	0,513	0,753	0,916	0,827	0,460	0,734	0,786	Etkilenmemiş
0,000	0,000	-	0,000	-	-	0,424	0,006	Ölü

4.2.5. Karışıklık matrisi

Bir karışıklık matrisi, tahmin edilen sınıf sayısının N olması durumunda N*N lik bir matrisi temsil etmektedir.

Bizim durumumuz için karışıklık matrisinin, yaralı sayılarını, etkilenmeme ve ölüm sayılarını belgelememize izin verdiğini belirtmek isteriz.

- Yaralı sayısını
- Etkilenmemiş kişi sayısını
- Ölüm sayısını

temsil etmektedir.

Tablo 4.6. Karışıklık matrisi

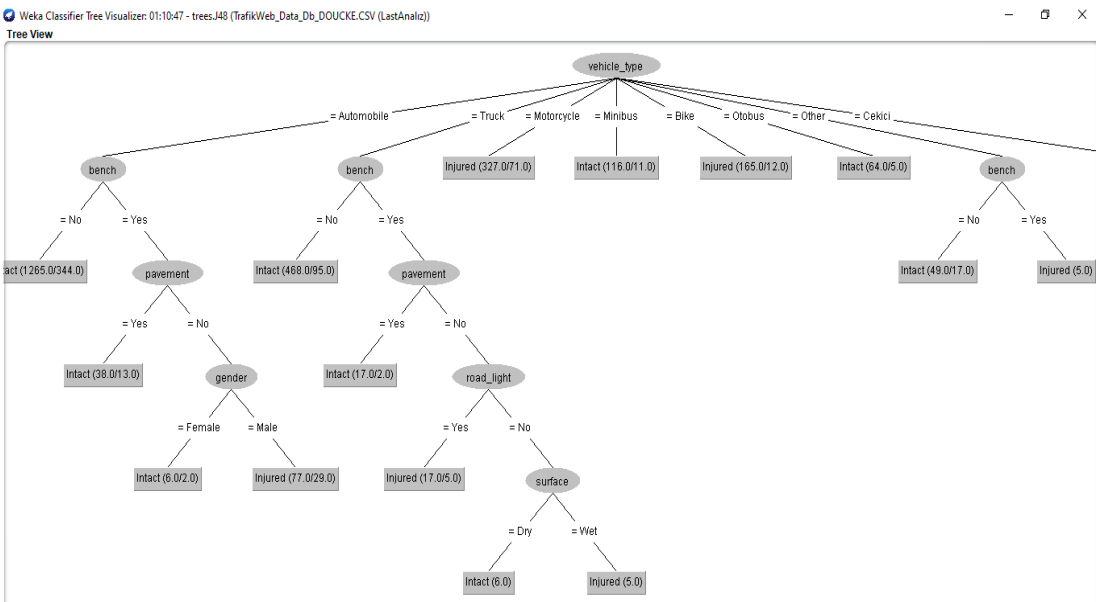
	A	B	C
A	470	492	0
B	139	1520	0
C	3	7	0

4.2.6. Karar ağacının grafiği

WEKA'da J48 karar ağacı algoritmasını kullanmak için aşağıdaki adımlar uygulanmıştır.

- Keşfet'e tıkla
- Seç'e tıkla
- Dosyaya tıkla
- Sınıflandır'a tıkla
- Seç'e tıkla
- Ağaçlara tıkla
- J48'e tıkla
- Değişken niteliğinin adına tıkla (Sonuç)
- Başlat'a tıkla
- Ağaçları görselleştir'e tıkla

J48 algoritması çalıştırıldıktan sonra elde edilen karar ağacı ile ilgili görsel aşağıdaki şekilde gösterilmiştir.



Şekil 4.4. Karar ağacı grafiği

4.2.7. Karar ağacının yorumlanması

Araç Tipi olan Düğüm 1, motosiklet sürücülerinin toplamda karıştıkları 327 kayıtlı kazada 71 motosiklet kullanıcısının yaralanmaya maruz kaldığı görülmektedir. Oransal olarak bakıldığında motosiklet kullanıcılarının diğer araç kullanıcılarına göre daha yüksek yaralanma oranlarına sahip oldukları görülmektedir. Araç tipine ait olan düğüm otomobil, kamyon ve diğer araç tipleri olmak üzere üç alt düğüme bölünerek, bu araç tipleri açısından kazaların şekillenmesi ve etkilenmesi ile ilgili veri ilişkileri belirlenmiştir.

Araç tipi düğümün otomotiv kökü tarafından sürülen bank düğümü yarıya indirilir. Hayır olduğunda ise 1265 kaza oranının motorlu araç tipi sürücüler tarafından gerçekleştiğini görüyoruz. Böylece kaydedilen 1265 kaza oranından sadece 344 araba kazadan sonra sağlam kaldı. Şimdi kökü evet denilince bizi kaldırım düğümüne götürüyor.

Düğüm seviyesinde, kaza üstyapıda meydana geldiğinde sadece 13 otomobilin sağlam olması nedeniyle, üstyapı düğüm seviyesinde üretilen kaza oranı 38 kaydedilmiştir. Bundan sonra kaldırım düğümünün kökünde ne zaman hayır dendiğini görüyoruz, bu bizi cinsiyet düğümüne götürüyor.

Cinsiyet düğümü düzeyinde, cinsiyet düğümünün ikiye, yani kadın ve erkek olarak ayrıldığını görüyoruz. Kadın ve erkek sürücüler arasında, otomobil kullanan kadın sürücülerin erkek sürücülerden daha az kaza yaptığını bulduk.

Düğüm sırası, tipik araç düğümünden gelen kamyon kökü tarafından sürülür. Kazaya uğrayan kamyonların oranı 468 olarak kaydedildi. O zaman sadece 95 kamyon sağlamdı. Yani kamyon düğüm bankının köküne hayır denildiğinde bizi kaldırım düğümüne götürür.

Kaldırım seviyesinde, kaldırımda meydana gelen kazanın meydana geldiği kamyon sayısından 17'sinin ve kaldırım seviyesinde kamyon şoförlerinin neden olduğu

kazadan sonra sadece 2'sinin sağlam olduğunu görüyoruz. Kaldırım seviyesinde, kaldırımda meydana gelen kazanın meydana geldiği kamyon sayısından 17'sinin ve kaldırım seviyesinde kamyon şoförlerinin neden olduğu kazadan sonra sadece 2'sinin sağlam olduğunu görüyoruz. Kamyon kaldırım düğümünün köküne hayır denildiğinde, bu bizi yol ışık düğümüne götürür. Yol ışık düğümü seviyesinde, 17 kaza vakasının kaydedildiğini ve yüzde 5'inin ciddi kazalar yaşadığını görüyoruz. Çünkü bazı kamyon şoförleri trafik ışıklarına uymadı. Ağacımızdan sonra yüzey düğümüne yol açtı.

Araç tipi düğümü bizim kökümüzdür. Dolayısıyla karar ağacına ilişkin yorumumuzu buradan yaparız. Araç tipi otomobil, bench ve asfalt (pavement) da evet olduğunda, tüm kazaların 38 tanesi yaralanmalı, 13 tanesi ise yaralanmasız olarak gerçekleşmiştir. Araç (vehicle) tipi kamyon (truck), bench evet, asfalt (pavement) hayır, ışıklandırma (road light) hayır, yol yüzeyi ise (road surface) evet olduğunda sadece 5 yaralanmalı kaza gerçekleşmiştir.

Araç tipi motorsiklet (motocycle) olduğunda toplam 327 yaralanmalı kaza gerçekleşirken, 71 adet de yaralanmasız kaza gerçekleşmiştir. Araç tipi minibüs olduğunda 116 hasarsız ve 11 hasarlı kaza belirlenmiştir. Araç tipi bisiklet (Bike) olduğunda 165 adet yaralanmalı, 12 adet de yaralanmasız kaza gerçekleşmiştir. Araç tipi otobüs olduğunda ise, 26 hasarsız, 5 hasarlı kaza ortaya çıkmıştır.

Yüzey düğüm seviyesinde, motorsiklet sürücülerinin yol ıslakken çok daha yüksek oranda kaza yaptığı analiz sonucunda ortaya çıkmıştır. Bununla beraber, birkaç ciddi kazanın yol yüzeyinin kuru olduğu durumda gerçekleştiği belirtilmelidir.

BÖLÜM 5. SONUÇ

5.1. Sonuç

Bu araştırma, kentsel alanlarda sürücülerin karıştığı kazaların analizine ve özellikle yol ağı uzmanlarını kolaylaştırmak için iyi görselleştirme ve kolay yorumlamayı kolaylaştırmak için verileri bir karar ağacının altında tutmanın doğru yoluna odaklandı. 2911'den fazla veriyi excel ‘de görselleştirmek zor olduğu için, karar ağacı ile verileri görselleştirmek ve bir şehirde veya bir ülkedeki kaza nedenleri hakkında iyi bir analiz yapmak daha kolaydır. Trafik kazaları, ölüme ve ciddi yaralanmalara neden olan küresel bir sorun olarak görülmektedir. Trafik kazalarında meydana gelen yaralanmaların seviyesinden risk faktörleri veya trafik kazası riskini artırabilecek diğer faktörler sorumludur. Bu çalışmada, yaralanma düzeyi ile ilgili 19 potansiyel gizli faktör seçilmiştir. Bu çalışmada Türkiye'nin Sakarya ilinde meydana gelen trafik kazaları incelenmiştir. Bu çalışmanın verileri Sakarya İl Emniyet Müdürlüğü'nden elde edilmiştir. Veri setinde 35 adet engelleme özelliği bulunmaktadır. 19 bağımsız değişken ve bağımlı değişken arasındaki ilişkiyi anlamak için, diğer veri madenciliği tekniklerine kıyasla uygulanan modelin net bir şekilde yorumlanması da dahil olmak üzere birçok avantajı olduğu için J48 algoritmasını kullanan sınıflandırma yöntemi seçilmiştir.

Karar ağaçlarından elde edilen sonuca göre, motosiklet sürücülerinin yaralanma oranı diğer ulaşım araçlarına göre daha yüksektir. Kamyon sürücüleri için, araba kuru olduğunda daha az kaza olduğunu fark ettik.

Analiz çalışmaları, motosiklet sürücülerinin karıştıkları kaza oranlarının oldukça yüksek olduğunu ortaya koymuştur. Dolayısı ile bu sistem kullanıcılarının eğitimi ve denetimleri ile ilgili olarak daha sıkı ve ciddi süreçlerin işletilmesi gerekmektedir.

Karar alma süreçlerinin etkin ve hızlı bir yapıda devreye sokulabilmesi adına WEKA oldukça net ve kolay anlaşılabilir karar ağacı analizleri ile, çok sayıdaki karmaşık verinin analizini yapabilmekte ve sağlıklı stratejilerin geliştirilmesine katkı sunabilmektedir.

KAYNAKLAR

[1] A. Y. Ng and M. I. Jordan, 'On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes', in Advances in neural information processing systems, 2002, pp. 841–848.

[2] T. Yamamoto, J. Hashiji, and V. N. Shankar, 'Underreporting in traffic accident data, bias in parameters and the structure of injury severity models', Accident Analysis & Prevention, vol. 40, no. 4, pp. 1320–1329, 2008.

[3] L. Li, C. G. Prato, and Y. Wang, 'Ranking contributors to traffic crashes on mountainous freeways from an incomplete dataset: A sequential approach of multivariate imputation by chained equations and random forest classifier', Accident Analysis & Prevention, vol. 146, p. 105744, 2020.

[4] T. B. Tesema, A. Abraham, and C. Grosan, 'Rule mining and classification of road traffic accidents using adaptive regression trees', International Journal of Simulation, vol. 6, no. 10, pp. 80–94, 2005.

[5] S. S. Alavi, M. R. Mohammadi, H. Souri, S. M. Kalhori, F. Jannatifard, and G. Sepahbodi, 'Personality, driving behavior and mental disorders factors as predictors of road traffic accidents based on logistic regression', Iranian journal of medical sciences, vol. 42, no. 1, p. 24, 2017.

[6] B. H. Ang, W. S. Chen, and S. W. H. Lee, 'Global burden of road traffic accidents in older adults: a systematic review and meta-regression analysis', Archives of gerontology and geriatrics, vol. 72, pp. 32–38, 2017.

[7] F. Ye and D. Lord, 'Investigation of effects of underreporting crash data on three commonly used traffic crash severity models: multinomial logit, ordered probit, and mixed logit', Transportation Research Record, vol. 2241, no. 1, pp. 51–58, 2011.

[8] J. C. Milton, V. N. Shankar, and F. L. Mannering, 'Highway accident severities and the mixed logit model: an exploratory empirical analysis', Accident Analysis & Prevention, vol. 40, no. 1, pp. 260–266, 2008.

[9] B. Dong, X. Ma, F. Chen, and S. Chen, 'Investigating the differences of single-vehicle and multivehicle accident probability using mixed logit model', Journal of Advanced Transportation, vol. 2018, 2018.

[10] M. Rezapour, S. S. Wulff, and K. Ksaibati, ‘Examination of the severity of two-lane highway traffic barrier crashes using the mixed logit model’, Journal of safety research, vol. 70, pp. 223–232, 2019.

[11] F. Chen, S. Chen, and X. Ma, ‘Analysis of hourly crash likelihood using unbalanced panel data mixed logit model and real-time driving environmental big data’, Journal of safety research, vol. 65, pp. 153–159, 2018.

[12] S. Bekhor, T. Toledo, and L. Reznikova, ‘A path-based algorithm for the cross-nested logit stochastic user equilibrium traffic assignment’, Computer-Aided Civil and Infrastructure Engineering, vol. 24, no. 1, pp. 15–25, 2009.

[13] W. Alajali, W. Zhou, S. Wen, and Y. Wang, ‘Intersection traffic prediction using decision tree models’, Symmetry, vol. 10, no. 9, p. 386, 2018.

[14] D. Shinar, Safety and mobility of vulnerable road users: pedestrians, bicyclists, and motorcyclists. Elsevier, 2012.

[15] Estimate road traffic death (per 10000 population). http: who.int/data/gho/indicators-details/GHO/estimated-road-traffic-death-rate.(per-100-000-population)(accessed May 05, 2021).

[16] National Hightway Traffic Safety Administration. 2016 Fatal motor vehicule craches: overview. Washington, DC : US Departement of transportation; 2017, W475W9360

[17] Lebigdata.fr/machine-learning-et-data/Data Analytics/Bastien L/fevrier 2021

[18] Build a Decision Tree in Minutes using Weka (No Coding Required!)ANIRUDDHA BHANDARI, MARCH 10, 2020

[19] Comment devenir Microsoft Certified Data Analyst Associate ?29 juillet 2021

[20] Machinelearningmasteery.com/ use-classification-machine-algorithms-Weka, Jason Brownlee June 2021

[21] IJISET-International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, February 2015.

ÖZGEÇMİŞ

Adı Soyadı **:** Joseph Doucke OMENDE KAHUDI

ÖĞRENİM DURUMU

Derece	**Eğitim Birimi**	**Mezuniyet Yılı**
Yüksek Lisans	Sakarya Üniversitesi / Fen Bilimleri Enstitüsü/ İnşaat Mühendisliği	Devam Ediyor
Lisans	Institut Du Batiment Et Travaux Publics Kinshasa/ Ngaliema	2015
Lise	Complexe Scolaire Cardinal Malulu Kinshasa Limite	2009

İŞ DENEYİMİ

Yıl	**Yer**	**Görev**
2014-2015	Ministere dlnfrastructure, travux publics et reconstruction da le RDC	Proje asistanı

YABANCI DİL

İngilizce, Fransızca, Türkçe

HOBİLER

Basketbol Oynamak