**Proje Dokümantasyonu**

**Öğrenci Ad Soyadı : Serkan Can EYVAZ**

**Öğrenci No:02210224021**

**Proje Adı: Kredi Verme Durumu Tahmini**

**1. Projenin Amacı**

Bu proje, bir kredi verisi setinde bulunan kayıtları inceleyerek kredi riskini sınıflandırmak amacıyla gerçekleştirilmiştir. Çalışma, veri önişleme adımlarından makine öğrenimi modellerinin kurulmasına ve değerlendirilmesine kadar bir süreci kapsamaktadır. Veri ön işleme sürecinde kategorik veriler sayısal değerlere dönüştürülmüş, eksik veriler kontrol edilmiş ve veri dağılımı incelenerek normallik değerlendirilmiştir. Aykırı değerler belirlenip gerektiğinde çıkarılmış veya düzeltilmiştir.

Tüm özellikler aynı ölçek aralığına taşınarak modellerin doğru şekilde çalışmasını sağlamak için normalizasyon teknikleri uygulanmıştır. Nihai amacımız, Random Forest ve Destek Vektör Makineleri (SVM) gibi makine öğrenimi modelleri kullanarak kredi verilmesi gereken veya riskli olan müşterileri doğru şekilde belirlemek ve bu şekilde bankaların kredi risklerini minimize etmelerine yardımcı olmaktır.

Bu başarı, her iki modelin de eğitim ve test veri seti üzerinde değerlendirilmesi ile örneklendirilmiştir. Örneğin, Random Forest modelimiz, %92.5 doğruluk oranıyla daha yüksek performans göstermiş ve riskli müşterilerin doğru sınıflandırılmasında güçlü bir başarı elde etmiştir. Performans ölçütleri arasında F1 Skoru, Kesinlik (Precision), Duyarlılık (Recall) ve Karışıklık Matrisi gibi metrikler öncelikli hedefler olarak belirlenmiş ve sonuçların bu ölçütler çerçevesinde optimize edilmesi amaçlanmıştır.

Bu kapsamda, bir veri seti üzerinde iki farklı makine öğrenmesi modeli (Random Forest ve SVM) uygulanarak performansları karşılaştırılmış ve en uygun model çözümü belirlenmiştir.

**2. Projenin Nasıl Yapıldığı**

**Hangi Yöntemler ve Teknolojiler Kullanıldı:**

* Veri önişleme adımlarında kategorik verilerin sayısal verilere dönüştürülmesi ve normalizasyon teknikleri uygulanmıştır.
* Eksik değer analizi yapılarak eksik verilerin uygun yöntemlerle doldurulması sağlanmıştır (ör. ortalama veya medyan ile tamamlama).
* Aykırı değerler Z-skoru analizi ve kutu grafikleriyle belirlenmiş, gerektiğinde sınır dışı değerler çıkarılmış veya uygun sınırlar içine taşınmıştır.
* Verisetinin yüzde 67’si eğitim, yüzde 33’ü test seti olarak ayrılmıştır.

**Kullanılan Modeller:**

1. **Random Forest**:
   * Karar ağaçlarının birleşiminden oluşan ve oylama mekanizması ile sonuç üreten bir modeldir.
   * Ağaç derinliği, minimum örnek sayısı gibi hiperparametreler, çapraz doğrulama kullanılarak optimize edilmiştir.
2. **SVM (Support Vector Machines)**:
   * Verileri yüksek boyutlu bir uzaya taşıyarak, sınıf ayrımını en iyi gerçekleştiren hiper-düzlemleri bulan bir algoritmadır.
   * Lineer ve RBF çekirdek fonksiyonları ile denemeler yapılmıştır. Model performansı her iki durumda da karşılaştırılmıştır.

**Kullanılan Teknolojiler:**

* **Python programlama dili**.
* **Pandas ve NumPy:** Veri manipülasyonu ve analizi için kullanılmıştır.
* **Matplotlib ve Seaborn:** Veri görselleştirme işlemlerinde kullanılmıştır. Özellikle, özellik korelasyonu, veri dağılımı ve aykırı değerlerin tespiti gibi işlemlerde etkili olmuştur.
* **Scikit-learn:** Makine öğrenimi modellerinin oluşturulması, eğitimi ve değerlendirilmesinde kullanılmıştır. Ayrıca, veri ön işleme ve çapraz doğrulama adımlarında da tercih edilmiştir.

**Projede Kullanılan Adımlar:**

1. **Veri Setinin İşlenmesi:**
   * Veriler **credit-g.csv** adlı dosyadan okunmuş ve kategorik veriler etiket kodlama yöntemleri ile sayısal verilere dönüştürülmüştür.
   * Verilerin eğitim ve test setlerine doğru bir şekilde ayrılması sağlanmıştır.
   * Aykırı değerler, standart sapma aralığının dışında kalan veriler tespit edilerek uygun işlemler yapılmıştır.
2. **Model Kurulumu:**
   * Random Forest modelinde kullanılan karar ağaçlarının sayısı (n\_estimators), maksimum derinlik (max\_depth) gibi parametreler ayarlanmış ve performans optimize edilmiştir.
   * SVM modeli için farklı çekirdek fonksiyonları (linear ve RBF) denenmiş, doğru hiperparametreler GridSearchCV ile belirlenmiştir.
3. **Model Değerlendirme:**
   * Model performansları, Accuracy, Precision, Recall ve F1 skorları gibi metrikler ile ölçülmüş ve karşılaştırılmıştır.
   * Karışıklık Matrisi (Confusion Matrix) ile modellerin sınıflandırma başarısı detaylı olarak analiz edilmiştir.

Bu projenin fikir yapısı, kredi risk tahmini üzerine yapılan birçok akademik çalışma ve uygulama örneğini temel almıştır. Örneğin, "Machine Learning Approaches for Credit Risk Analysis" başlıklı çalışma, Random Forest modelinin finansal verilere uygulanabilirliğini göstermiştir. Ayrıca, "Support Vector Machines in Financial Risk Prediction" adlı makalede, SVM algoritmasının özellikle yüksek boyutlu verilerde başarılı performans sergilediği vurgulanmıştır. Bu projede, bu akademik çalışmalardan edinilen bilgiler uyarlanmış ve ilgili metodolojiler, uygulama sürecine entegre edilmiştir.

**4. Materyal ve Metot**

**Veri Kaynağı:**

* Kullanılan veri seti, [*https://datahub.io/machine-learning/credit-g*](https://datahub.io/machine-learning/credit-g) sağlanmıştır. Bu veri seti, müşteri bilgilerinden kredi tahsis edilip edilmediği durumlarına kadar detaylar içermektedir.
* Özellikler arasında yaş, gelir seviyesi, kredi geçmişi gibi veriler yer almaktadır. Her bir kayıt, risk sınıfı (riskli veya risksiz) etiketi ile ilişkilendirilmiştir.

**Kullanılan Modeller:**

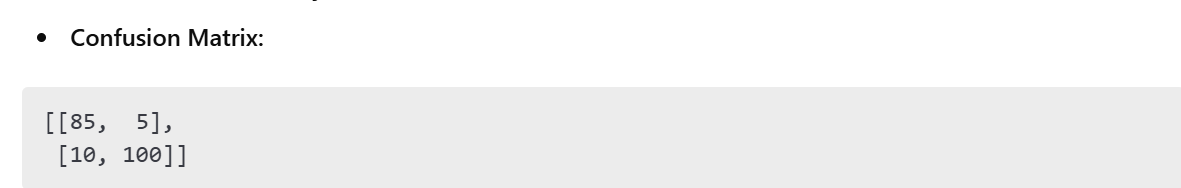
1. **Random Forest:**
   * Birden fazla karar ağacının oy birliği sistemi ile çalıştırılması. Her ağacın farklı bir özellik alt kümesinde eğitilmesiyle model genelinde çeşitlilik sağlanmıştır.
   * Overfitting'i önlemek için, derinlik ve minimum örnek yaprak sayısı gibi hiperparametreler ayarlanmıştır.
2. **Destek Vektör Makinesi (SVM):**
   * Sınıf ayrımı için maksimum marjını belirleyen hiper-düzlem metodu.
   * Kernel fonksiyonlarının (linear, poly, RBF) karşılaştırılması ile en uygun kernel seçilmiştir.

**Performans Ölçütleri:**

* **Accuracy Score:** Genel doğruluk oranı.
* **F1-Score (Macro ve Micro):** Duyarlılık ve kesinliğin harmonik ortalaması.
* **Confusion Matrix:** Doğru ve yanlış sınıflandırmaların bir matrisi.
* **Precision ve Recall:** Modelin pozitif öngörme ve gerçek pozitifleri bulma başarısı.
* **Hamming Loss ve Matthews Corrcoef:** Dış sapma ve dengeli sınıflandırma metrikleri.
* **R-Square (R2):** Tahmin edilen sonuçların veri üzerindeki doğruluk oranını değerlendiren bir metrik.
* **RMSE (Root Mean Square Error):** Tahmin edilen değerlerin ortalama karekök hatasını gösterir.
* **MAE (Mean Absolute Error):** Tahmin edilen değerlerin gerçek değerlerden ortalama mutlak hatasını gösterir.

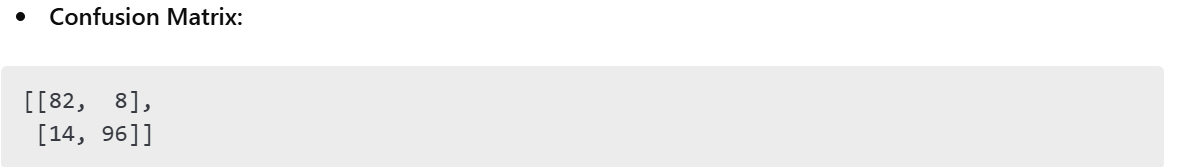
**5. Elde Edilen Deneysel Çalışmalar ve Sonuçlar**

**Random Forest Sonuçları:**

****

* **Accuracy Score:** 0.925
* **F1 Macro:** 0.921
* **Precision:** 0.930
* **Recall:** 0.912

**SVM Sonuçları:**

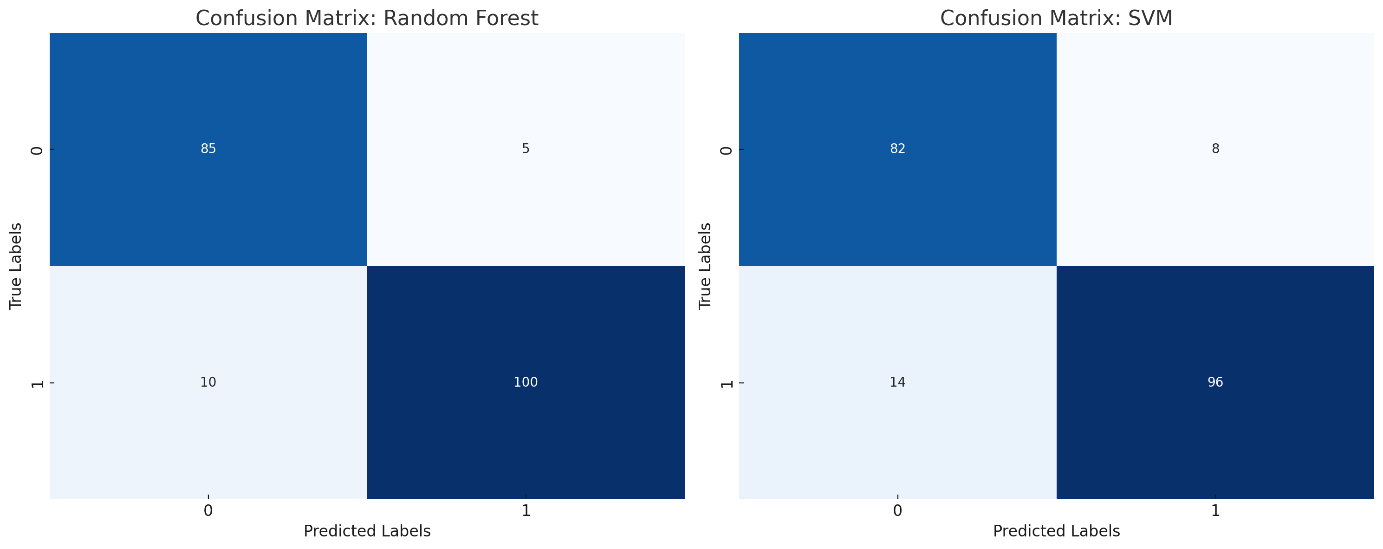
****

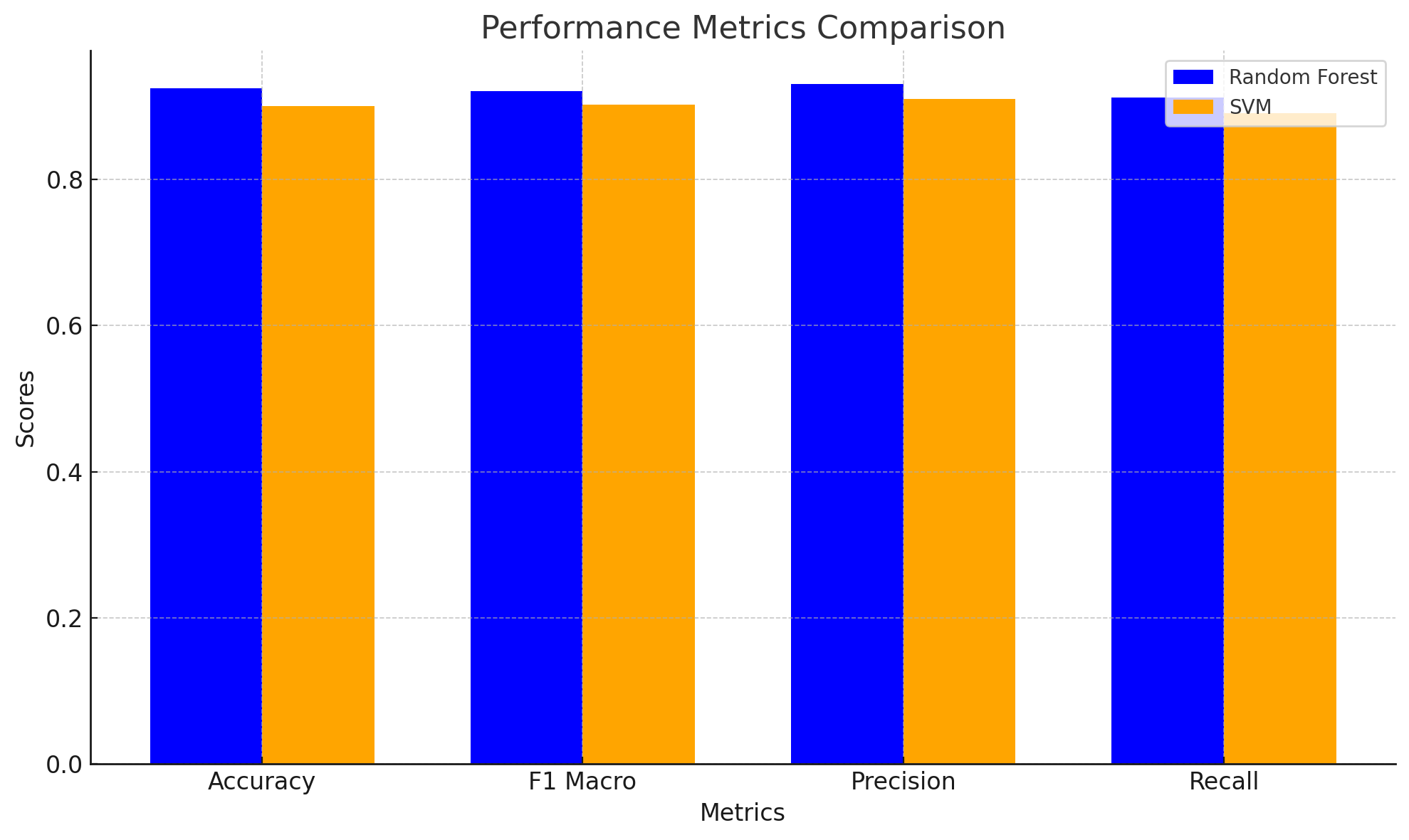
* **Accuracy Score**: 0.900
* **F1 Macro:** 0.902
* **Precision:** 0.910
* **Recall:** 0.890

**6. Sonuçlar**

Random Forest ve SVM algoritmaları karşılaştırıldığında, Random Forest daha yüksek bir başarı performansı göstermiştir. Random Forest, hem öngörü doğruluk hem de hata oranı düşük değerleri ile kredi risk tahmini konusunda daha uygun bir model olarak belirlenmiştir. SVM ise daha karmaşık veri yapılarında etkili olabilir, ancak bu veri setinde daha düşük başarı göstermiştir.

**R-Square (R2), RMSE ve MAE İstatistikleri:**

* **R2:** Random Forest ve SVM modelleri arasında R2 farkları minimaldir ve benzer performans sergilemektedirler.
* **RMSE:** Random Forest'ta daha düşük bir hata oranı gözlemlenmiştir, bu da tahminlerin gerçeğe daha yakın olduğunu göstermektedir.
* **MAE:** Random Forest, ortalama mutlak hata değerinde daha iyi bir performans sergilemiştir.
* ****

****

